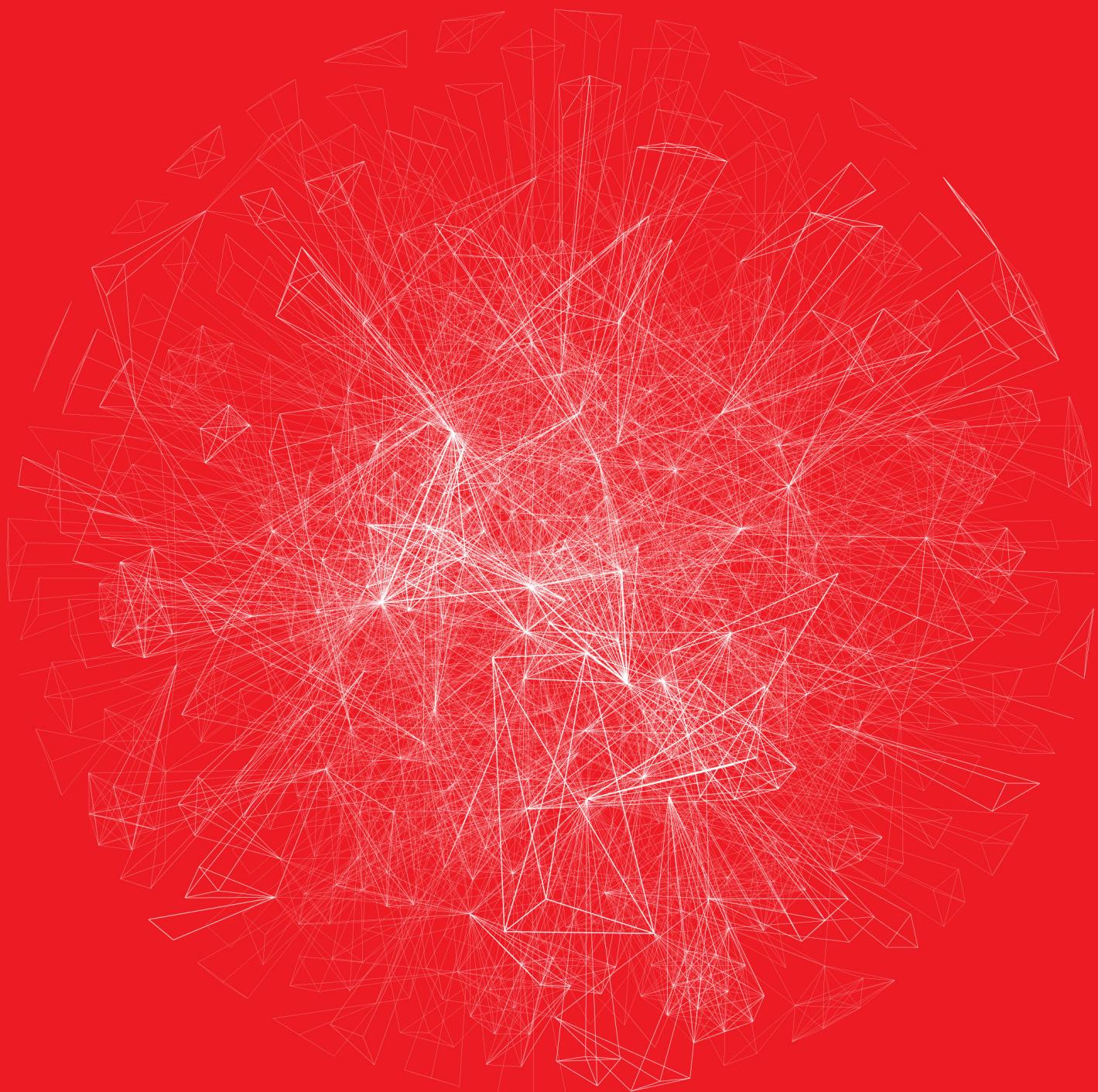


■ Digital Humanities 2014

Book of Abstracts



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 2014

Conference Abstracts
EPFL - UNIL
Lausanne, Switzerland
8-12 July 2014

Main Coordinator:

Cyril Bornet

Designers:

Dario Rodighiero

Victor Audéat

Encoders and Proofreaders:

Maude Auderset

Kevin P. Baumer

Vincent Buntinx

TEI to PDF tools:

Karin Dalziel

Available online at: <http://dh2014.org>

The 25th Joint International Conference of the Association for Literary
and Linguistic Computing and Association for Computers and the Humanities
and

The 6th Joint International Conference of the Alliance of Digital Humanities Organizations



International Program Committee

Melissa Terras, chair
Deb Verhoeven, vice-chair

John Bradley
Jieh Hsiang
Jane Hunter
Aimée Morrison
Bethany Nowviskie
Dan O'Donnell
Sarah Potvin
James Smithies
Takafumi Suzuki
Tomoji Tabata
Toru Tomabechi
Glen Worthey
Vika Zafrin

Local Organizing Committee

Claire Clivaz, co-local organizer
Frédéric Kaplan, co-local organizer

Karl Aberer
Jeannette Frey
Benoît Garbinato
Philippe Kaenel
Isabelle Kratz
Enrico Natale
Lukas Rosenthaler
Süsstrunk Sabine
Michael Stolz
François Vallotton
Boris Vejdovsky
Dominique Vinck

Silver Level Sponsors

Yandex Europe AG

Bronze Level Sponsors

CLARIN
Common Language Resources and Technology Infrastructure

Partners

infoclio.ch
Swiss National Science Foundation

Organizers

Alliance of Digital Humanities Organizations (ADHO)
École Polytechnique Fédérale de Lausanne (EPFL)
Université de Lausanne (UNIL)

Conference Volunteers

Cyril Bornet
Diane Brousse
Vincent Buntinx
Giovanni Colavizza
Olivier Dalang
Isabella di Lenardo
Heidi Dowding
Sebastien Dupont
Anthony Durity
Mikal Eckstrom
Maud Ehrmann
Slimane Fouad
Alicia Foucart
Hannah Jacobs
Penny Johnston
Andrea Mazzei
Elisa Nury
Paul O'Shea
Anna Pytlowany
Yannick Rochat
Jörg Röder
Dario Rodighiero
Elifsu Sabuncu
Jillian Saucier
Qiaoyu Shi
Sree Ganesh Thottempudi

Welcome to Digital Humanities 2014

Professor Frédéric Kaplan

frédéric.kaplan@epfl.ch
Co-local organizer, DHLab, EPFL, Lausanne

Professor Claire Clivaz

claire.clivaz@unil.ch
Co-local organizer, Ladhul, UNIL, Lausanne

We are proud to welcome the *Digital Humanities 2014* conference to Lausanne, which marks the first time it will be held in Switzerland. We are also proud because this event represents two milestones. Firstly, as a reflection of the recent development of digital humanities in Switzerland, and secondly as a crucial step in the long histories of our two co-hosting institutions: the Ecole Polytechnique Fédérale de Lausanne (EPFL) and The University of Lausanne (UNIL).

Swiss people are great at creating new interfaces. By the time Swiss clocks and automata were anticipating our digital future, implicitly reflecting on the multiple ways men and machines could coordinate their actions, Germaine de Staél was opening new intellectual paths in her reunions at the castle of Coppet; the European network was in preparation. More recently, ten years after a Swiss company – Logitech – transformed the “mouse” into an interface used worldwide, completely reshaping the gesture grammars of computer interactions, The World Wide Web, surely one of the most culturally significant interfaces ever imagined, was invented in Geneva. Since then, Switzerland has been on the leading edge of the cultural and technological developments underpinning the “digital revolution.” Therefore, it was natural for us to suggest “Digital Cultural Empowerment” as the theme for this year’s conference – encouraging participants to reflect on the way the digital era is reshaping the way we live, work and think, transforming all aspects of our diverse cultures.

As Switzerland is a multicultural and multilingual country, we wanted this conference to embrace this diversity. Thanks to the help of the MLCM committee, we are very pleased that this year’s call for papers was available in 23 languages, including Arabic and Chinese. This is a symbolic turning point for the digital humanities domain, still largely anchored in Western points of view. There is a common willingness to increase cultural and linguistic diversity in our community, as was recently exemplified by the creation of GO::DH (Global Outlook Digital Humanities). The recent launching of new linguistic digital humanities associations, such as the French-speaking and Spanish-speaking ones, is also a sign of the growing cultural diversity in digital humanities. This cultural opening brings the promise of great transformations in the domain, beyond the limit of Western-centric approaches and more towards globally interfaced communities of best practices.

As in many other countries, departments in Swiss universities, such as the Institut für Computerlinguistik in Zürich, or Imaging Medialab in Basel, were running projects at the crossroads of computing and humanities, long before the term digital humanities was used. The digital humanities label gained momentum in Switzerland in 2010, via discussions held between a small team of Unil and EPFL colleagues, who were convinced that research in this domain could offer great opportunities for development in the fields of Humanities and Computer Science. With the support of EPFL and Unil management, two Digital Humanities laboratories were opened.: the DHLAB at EPFL in July 2012 and Ladhul at Unil in January 2013. Simultaneously the University of Bern announced the opening of three assistant professor positions under the digital humanities banner, and the Imaging Medialab was renamed DHLAB Basel.

The digital humanities field is now recognized as strategically important for the future of research in Switzerland. In November 2013, it was the central theme of the annual meeting of the Swiss Academy for Humanities. Several large-scale digital humanities projects have now been launched. For example, The Venice Time Machine Project, which is a collaboration between EPFL, the University Ca’Foscari and the State Archive in Venice, which is building a historical simulation of Venice and its empire over a 1000 year period. There is also the CUS P2 program, providing a national strategy for an extensive range of science-related digital content and associated tools, and the DDZ/CDP project, dealing with long-term preservation and reuse of databases of humanities-related content, as well as several other digitization programs (Swiss press digitization program, the Bodmer Foundation digitization program, etc.). In this environment of rapid development, hosting *Digital Humanities 2014* is a crucial milestone in the short but lively adventure of digital humanities in Switzerland.

Digital Humanities 2014 is organized by two neighboring institutions, complementary in many respects and sharing the same large campus on the shore of Lake Geneva. This undertaking was a unique occasion for both teams to collaborate, combining their skills and knowledge towards one single goal: offering the best experience possible to our international guests. We sincerely thank all the key players from both institutions, who have worked tirelessly towards the success of this event, and wish that this exciting moment will play a key role in further strengthening the links between our two institutions.

Our warm thanks go to all the sponsors, both public and private, that believe in this event and have made it possible with their financial support - The Swiss National Science Foundation, Infoclio, Yandex, Open Edition – as well as to the many volunteers who have helped to make this event a success.

We really hope you enjoy *Digital Humanities 2014* and your time the city of Lausanne.

Welcome from the Program Committee Chair

Professor Melissa Terras

m.terras@ucl.ac.uk

Director, UCL Centre for Digital Humanities, University College London

It is my pleasure to present this bumper edition of conference abstracts for DH2014. This is the 26th joint annual conference of (what is now) the European Association for Digital Humanities and the Association for Computers and the Humanities, joined together under the umbrella of the Alliance of Digital Humanities Organisations, which also is now constituted with the Canadian Society for Digital Humanities (CSDH/SCHN), the Australasian Association for Digital Humanities (aaDH), the Japanese Association for Digital Humanities (JADH), and centerNet. Our organisations and membership continue to grow, and this year our pre-conference meetings also host the first annual meeting of the French-speaking Digital Humanities community: Humanistica.

The planning for a conference as large as this generally begins eighteen months in advance, and this book of abstract sees the end of a long and rigorous process of organisation, publicising, reviewing, choosing, and planning the contents you now see presented here. This year the call for papers was issued in 27 different languages, and we had submissions primarily in English, but also in Arabic, German, French, Portuguese, Spanish, Italian and a paper with significant Japanese content: our field is truly international.

Including workshops and bursaries there were over 750 submissions on confitool, our conference management system, with 600 proposals submitted for the first phase of papers, panels, and posters, from over 2000 submitting authors. We are pleased to be able to open up another parallel track to accommodate more presentations by members of our growing community, and for this the Program Committee would like to thank our local hosts at EPFL and Unil for accommodating the plea for extra space, to showcase the work you see here.

The acceptance rate for panels, long papers and short papers was just under 50%, with many more posters slots being made available this year, bringing the overall acceptance rate to just under 60%. 40 workshops were submitted, of which 28 were accepted across a range of training and discussion areas. This means that DH2014 is the largest ever Digital Humanities conference in the history of the joint annual meeting, and the fierce competition for places indicates the growth and maturing of our field.

Managing such a process depends on the goodwill and hard work of both our Program Committee, and our relatively small band of peer reviewers, who pulled together to ensure each and every submission had an adequate number of reviews to make this process as fair as possible. A list of reviewers is presented later in this volume, and we are indebted to the hard work of these individual scholars: as our community continues to grow at a rapid pace, we need to ensure we can accommodate the level of peer review required to uphold academic standards, and the efforts of some peer reviewers to help out in this year of such interest in the conference should not be underestimated (I am particularly grateful to one individual who undertook twenty mostly panel reviews!). Thank you for all of your contributions to the process: this truly is a program the community made.

The program committee have been my crew, toiling away quietly behind the scenes, and my appreciation goes out to them all: the Vice Chair: Deb Verhoeven (aaDH), John Bradley (EADH), Jieh Hsiang (Centernet), Jane Hunter (aaDH), Aimée Morrison (CSDH/SCHN), Dan O'Donnell (CSDH/SCHN), Sarah Potvin (Centernet), James Smithies (aaDH), Takafumi Suzuki (JADH), Tomoji Tabata (EADH), Toru Tomabechi (JADH), Glen Worthey (ACH), Vika Zafrin (ACH). Last year's chair, Bethany Nowviskie (ACH), has been a wonderful act to follow, and I thank her for her sound advice and guidance on the peer reviewing process. Many thanks are also due to John Nerbonne, who has been leading ADHO's Conference Coordinating Committee, and Neil Fraistat, who has chaired the ADHO Steering Committee throughout the planning phase for DH2014. The members of both of those committees also responded to many pleas, questions, and points of process inevitably raised in the planning of DH2014. The Call for Papers was translated in a phase led by Elizabeth Burr, chair of the Multilingual and Multicultural Issues Committee, and we thank all of those who gave their time for this task.

As I write this introduction, the phase of program creation is all but over, but sincerest thanks are due to the Local Organisers, Claire Clivaz (Unil) and Frederic Kaplan (EPFL) and their team members, including Kevin Baumer and Cyril Bornet, for helping making this plan into an actual, physical conference and dealing with all logistical matters with efficiency and good cheer. I would also like to thank the 66 session chairs, who will help us both in keeping to our program, and also in building the intellectual framework on which this conference sits.

It has been an honour to be asked to lead this process, and to interact with so many of the digital humanities community over the past eighteen months as we have organised Digital Humanities 2014. Our Local Organisers pitched the theme "Digital Cultural Empowerment", and the range and breadth of submissions presented here demonstrates the complexity and scope of our activities across the Digital Humanities community. I wish you all a fruitful and intellectually stimulating conference, and am proud of the combined efforts of our community to produce such a varied, multi-faceted program.

Bursary Winners

Digital Humanities 2014 Student Conference Bursaries

Kawase, Akihiro (National Institute for Japanese Language and Linguistics, Japan)
Vitale, Valeria (King's College London, United Kingdom)
O'Sullivan, James Christopher (Pennsylvania State University / University College Cork, Ireland)
Paquette-Bigras, Ève (Université de Montréal, Canada)
Jänicke, Stefan (Leipzig University, Germany)
Dye, Dotty J (Arizona State University, France)
Gutiérrez De la Torre, Silvia Eunice (Würzburg Universität, Germany)
Hidalgo Urbaneja, María Isabel (Universidad de Málaga, Spain)
Hamilton, Rachael Louise (University of Glasgow, United Kingdom)
Grue, Dustin Elias (The University of British Columbia, Canada)
Swafford, Joanna Elizabeth (University of Virginia, United States of America)
Masotti, Raffaele; Kenny, Julia; Di Pietro, Chiara (Università di Pisa, Italy)
Peaker, Alicia Rose (Northeastern University, United States of America)
Gawley, James O'Brien; Forstall, Chris Walton (University at Buffalo, United States of America)
Trettien, Whitney (HyperStudio, MIT, United States of America)
Reeve, Jonathan Pearce (New York University, United States of America)
Plasek, Aaron Louis (New York University, United States of America)
Christie, Alexander (University of Victoria, Canada)

Table of Contents

List of Reviewers	1
-------------------------	---

Pre-Conference Workshops and Tutorials

Are we there yet? Functionalities, synergies and pitfalls of major digital humanities infrastructures	
Benardou, Agiatis; Champion, Erik; Hughes, Lorna; Chambers, Sally; Dallas, Costis; Dunning, Alastair	4
Introducing the EpiDoc Collaborative: TEI XML and tools for encoding classical source texts	
Bodard, Gabriel; Franzini, Greta; Stoyanova, Simona; Tupman, Charlotte	5
Building bridges between Lausanne and Leeds: Virtual Round Table Discussion on methods, recent solutions and new questions between scholars at the International Mediaeval Congress in Leeds and the Digital Humanities Congress in Lausanne	
Bruhn, Kai-Christian; Schwartz, Frithjof	6
A Collaborative, Indeterministic and partly Automatized Approach to Text Annotation	
Bögel, Thomas; Gius, Evelyn; Petris, Marco; Strötgen, Jannik	7
Sharing digital arts and humanities knowledge: DARIAH as an open space for dialogue	
Chambers, Sally; Schmunk, Stefan	8
Hacking with the TEI	
Ciula, Arianna; Czmiel, Alexander; Mylonas, Elli; Rahtz, Sebastian; Cummings, James; Syd, Bauman	10
“What’s your method?” Building an ontology for digital research methods in the arts and humanities	
Constantopoulos, Panos; Dallas, Costis; Hughes , Lorna; Thaller, Manfred	11
Annotation Studio: an open-source, collaborative multimedia online note-taking tool for humanities teaching and learning	
Fendt, Kurt; Folsom, Jamie; Schnepper, Rachel; Andrew, Liam	12
GIS in the Digital Humanities: An introductory workshop	
Gregory, Ian; Barker, Elton; Lang, Anouk	14
Methods for Empowering Library Staff through Digital Humanities Skills	
Hettel, Jacqueline; Lindblad, Purdom; Baker, James; Stack, Padraig; Gil, Alex; Miller, Laura; Bourg, Chris	15
Introduction to Text Analysis and Topic Modeling with R	
Jockers, Matthew	17
Project management and sustainable revenue models in the Digital Humanities	
Keller, Stefan Andreas; Keller, Alice; Neuroth, Heike; Rosenthaler, Lukas	18
Linked Data and Literature: Encoding the Facts in Fiction	
Lawrence, Katharine Faith	19
Ontologies for Prosopography: Who’s Who? or, Who was Who?	
Lawrence, Katharine Faith; Bodard, Gabriel; Bradley, John; Perdue, Susan; Rahtz, Sebastian; Daniel, Pitti; Christian-Emil, Ore	20
The Representation of Multiplicity as a Means to Digital Cultural Empowerment	
Mareike, Hoeckendorff; Vitale, Valeria; Dunn, Stuart; Gius, Evelyn	22
Sound and (moving) images in focus – How to integrate audiovisual material in Digital Humanities research	
Ordelman, Roeland; Kemman, Max; Kleppe, Martijn; de Jong, Franciska	24
Digital Cultural Empowerment	
Palm, Fredrik; Murphy, Orla; Day, Shawn; Thély, Nicholas	26
Prosopography Workshop	
Quamen, Harvey; Crompton, Constance; Hjartarson, Paul	27
Leveraging Web Archiving Tools for Digital Humanities Research and Digital Exhibition	
Reed, Scott Brian	28
Multilinguality in historical documents – challenges and solutions for digital humanities	
Romary, Laurent; Dipper, Stefanie; Bubenhofer, Noah; Vertan, Cristina	29
DARIAH-EU VCC2 Workshop on Innovative Teaching Methods and Practices in Digital Humanities	
Scholger, Walter; Clivaz, Claire; Tasovac, Toma	31
Introduction to Starting and Sustaining DH Centers	
Siemens, Lynne	32
Kickstarting the GO:DH Minimal Computing Working Group	
Simpson, John Edward; Sayers, Jentery; O'Donnell, Daniel Paul; Gil, Alex	32
My Very Own Voyant: From Web to Desktop Application	
Sinclair, Stéfan; Rockwell, Geoffrey	34
Introduction to electronic books and EPub 3.0	
Sperberg-McQueen, Michael	35
Using the PressForward Plugin to Create and Maintain Web Publications	
Westcott, Stephanie; Fragaszy Troyano, Joan	36
Using CLARIN for Digital Research	
Wynne, Martin; Trippel, Thorsten; Draxler, Christoph	37
Curation, Management, and Analysis of Highly Connected Data in the Humanities	
de la Rosa, Javier; Brown, David Michael	38
Panels	
Annotating in Digital Music Edition - concepts, processes and visualisation of annotations	
Beer, Nikolaos; Bohl, Benjamin W.; Seuffert, Janette	42

<audio>Digital Humanities</audio>: The Intersections of Sound and Method	
Clement, Tanya; Kraus, Kari; Sayers, Jentery; Trettien, Whitney; Tcheng, David; Auvil, Loretta; Borries, Tony; Wu, Min; Oard, Doug; Hajj-Ahmad, Adi; Su, Hui; Lingold, Mary Caton; Mueller, Daren; Turkel, William J.; Elliott, Devon	44
New and recent developments in image analysis: theory and practice	
Crowther, Charles; Nyhan, Julianne; Tarte, Segolene; Dahl, Jacob	46
Remediating 20th-Century Magazines of the Arts: Approaches, Methods, Possibilities	
Ermolaev, Natalia; Wulfman, Clifford E.; Biber, Hanno; Crombez, Thomas	50
Global Outlook::Digital Humanities: Promoting Digital Humanities Research Across disciplines, regions, and cultures	
O'Donnell, Daniel Paul; Bordalejo, Barbara; Risam, Roopika; Spence, Paul; González-Blanco, Elena	54
Spectacle vivant et technologie numérique: du laboratoire scientifique au plateau de théâtre	
Pluta, Izabella; Fourmentraux, Jean-Paul; Bardiot, Clarisse	58
Rethinking Text Reuse as Digital Classicists	
Romanello, Matteo; Berra, Aurélien; Trachsel, Alexandra	62
What is Modeling and What is Not?	
Van Zundert, Joris; Jannidis, Fotis; Drucker, Johanna; Rockwell, Geoffrey; Underwood, Ted; Kestemont, Mike; Andrews, Tara	63

Papers

"Civilization arranged in chronological strata": A digital approach to the English semantic space	
Alexander, Marc; Anderson, Wendy	67
Metaphor, Popular Science and Semantic Tagging: Distant Reading with the Historical Thesaurus of English	
Alexander, Marc; Anderson, Jean; Baron, Alistair; Dallachy, Fraser; Kay, Christian; Piao, Scott; Rayson, Paul	68
The Cryptic Novel: A Computational Taxonomy of the Eighteenth-Century Literary Field	
Algee-Hewitt, Mark; Eidem, Laura; Heuser, Ryan; Law, Anita; Llewellyn, Tanya	70
The Stanford Literary Lab Transhistorical Poetry Project Phase II: Metrical Form	
Algee-Hewitt, Mark; Heuser, Ryan; Kraxenberger, Maria; Porter, J.D.; Sensenbaugh, Jonny; Tackett, Justin	72
Common Container Correlation: A Simple Method for the Extraction of Structural Models from Statistical Data	
Alvarado, Rafael	73
Rethinking Recommendations: Digital Tools for Art Discovery	
Andrew, Liam; Gonzalez, Desi	73
L'édition numérique – système d'organisation des connaissances avec les outils sémantiques	
Andréys, Clémence; Borel, Clément; Roxin, Ioan	75
From the Archimedes' Palimpsest to the Vercelli Book: Dual Correlation Pattern Recognition and Probabilistic Network Approaches to Paleography in Damaged Manuscripts	
Anthony, Eleanor Chamberlain	76
Tracing Workflow of a Digital Scholar	
Antonijevic, Smiljana; Stern Cahoy, Ellysa	76
Building a multi-dimensional space for the analysis of European Integration Treaties. An XML-TEI scenario	
Armaselu, Florentina; Allemand, Frédéric	77
The Layered Text. From Textual Zoom, Text Network Analysis and Text Summarisation to a Layered Interpretation of Meaning	
Armaselu, Florentina	80
Binarization-free Text Line Extraction for Historical Manuscripts	
Arvanitopoulos Darginis, Nikolaos; Süsstrunk, Sabine	83
Leaves of Grass: Data Animation and XML Technologies	
Barney, Brett; Pytlík Zillig, Brian	85
CURIOS: Connecting and Empowering Community Heritage through Linked Data	
Beel, David; Webster, Gemma; Mellish, Chris; Wallace, Claire	86
Unhappy? There's an App for That: Digital Happiness, Data Mining, and Networks of Well-Being	
Belli, Jill	87
Where is my Other Half?	
Ben-Shalom, Adiel; Choueka, Yaacov; Dershowitz, Nachum; Shweka, Roni; Wolf, Lior	89
Marrying the Benefits of Print and Digital: Algorithmically Selecting Context for a Key Word	
Benner, Drayton Callen	91
Riorganizzare SignWriting per favorire la ricerca linguistica sulle Lingue dei Segni	
Bianchini, Claudia S.; Borgia, Fabrizio; De Marsico, Maria	94
Uncertain about Uncertainty: Different ways of processing fuzziness in digital humanities data	
Binder, Frank; Entrup, Bastian; Schiller, Ines; Lobin, Henning	95
Mining a 'Trove': Modelling a Transnational Literary Culture	
Bode, Katherine	98
Distant reading of naïve poetry: corpora comparison as research methodology	
Bonch-Osmolovskaya, Anastasia; Orekhov, Boris	100
Dimensions of literary appreciation. Word use and ratings on a book discussion site	
Boot, Peter	102
An XML Schema to Interpret Networked Biographies: Reading Mid-Range	
Booth, Alison; Martin, Worthy	103
Scholarly primitives revisited: towards a practical taxonomy of digital humanities research activities and objects	
Borek, Luise; Dombrowski, Quinn; Munson, Matthew; Perkins, Jody; Schöch, Christof	105
Single Page Apps for Humanists: A Case Study using the Perseus Richmond Times Corpus	
Borg, Trevor; Thiruvathukal, George Kuriakose	107
A top-down approach to the design of components for the philological domain	
Boschetti, Federico; Del Grosso, Angelo Mario; Khan, Anas Fahad; Lamé, Marion; Nahli, Ouafae	109
Exploring a model for the semantics of Medieval Legal Charters	
Bradley, John; Rio, Alice; Hammond, Matthew; Broun, Dauvit	111

An XML annotation schema for speech, thought and writing representation	
Brunner, Annelen	112
Europe as a Digital Network: EGO European History Online	
Burch, Thomas; Berger, Joachim	114
Socially-Derived Linking and Data Sharing within a Virtual Laboratory for the Humanities	
Burrows, Toby; Verhoeven, Deb; Hawker, Alex	115
SyMoGIH project and Geo-Larhra: A method and a collaborative platform for a digital historical atlas	
Butez, Claire-Charlotte; Beretta, Francesco	116
Towards visualizing linguistic patterns of deliberation: a case study of the S21 arbitration	
Bögel, Tina; Gold, Valentin; Hautli-Janisz, Annette; Rohrdantz, Christian; Sulger, Sebastian; Butt, Miriam; Holzinger, Katharina; Keim, Daniel A.	117
Mining poetic rhythm: using text-to-speech software to rewrite English literary history	
Cade-Stewart, Michael	119
The Landscapes of Casta Paintings: Depictions of Social Anxieties in XVIII Century New Spanish Art	
Caldas, Natalia; Ortega, Érika; Jiménez Mavillard, Antonio; Brown, David; Suárez, Juan Luis	120
Matérialiser et rendre perceptible la transmission orale du savoir. L'édition électronique des cours d'Antoine Desgodets à l'Académie royale d'architecture en France, 1719-1728	
Carvais, Robert; Chateau, Emmanuel	123
Six terms fundamental to modelling transcription	
Caton, Paul	125
Z-Axis Scholarship: Modeling How Modernists Write the City	
Christie, Alex; Ross, Stephen; Sayers, Jentery; Tanigawa, Katie; INKE-MVP Research Team	126
Book History and Software Tools: Examining Typefaces for OCR Training in eMOP	
Christy, Matthew; Samuelson, Todd; Torabi, Katayoun; Tarpley, Bryan; Grumbach, Elizabeth	129
Diagnosing Page Image Problems with Post-OCR Triage for eMOP	
Christy, Matthew; Auvin, Loretta; Gutierrez-Osuna, Ricardo; Capitanu, Boris; Gupta, Anshul; Grumbach, Elizabeth	130
Developing for Distant Listening: Developing Computational Tools for Sound Analysis By Framing User Requirements within Critical Theories for Sound Studies	
Clement, Tanya	132
Beyond the Tool : A Reflexive Analysis on Building Things in Digital Humanities	
Couture, Stéphane; Sinclair, Stéfan	134
Validating Computational Stylistics in Literary Interpretation	
Craig, Hugh; Eder, Maciej; Jannidis, Fotis; Kestemont, Mike; Rybicki, Jan; Schöch, Christof	135
Readings of a photograph: Cognition and Access	
Das Gupta, Vinayak	137
The Scholarly 3D Toolkit: Annotation, Publication, and Analysis of 3D Scenes alongside Imported Humanities Data	
DataColtrain, James Joel	138
Digital Cultural Heritage and the Healing of a Nation: Digital Sudan	
Deegan, Marilyn	140
Mapping and Unmapping Joyce: Geoparsing Wandering Rocks	
Derven, Caleb; Teehan, Aja; Keating, John	141
DH on the Fringes: Using Smartphones, Instagram, and Ruby on Rails to Archive the DH Experience at an HBCU	
Dighton, Desiree; Norberg, Brian	143
Exploring the Intersection of Personal and Public Authorial Voice in the Works of Willa Cather	
Dimmit, Laura; Kirilloff, Gabrielle; Warren, Chandler; Wehrwein, James	144
Representation and Absence in Digital Resources: The Case of Europeana Newspapers	
Dunning , Alastair; Neudecker, Clemens	146
On Reusability and Electronic Literature	
Durity, Anthony; O'Sullivan, James	147
Digital Yoknapatawpha: Interpreting a Palimpsest of Place	
Dye, Dotty J.; Napolin, Julie Beth; Cornell, Elizabeth; Martin, Worthy	149
Digital Activism: Canon Expansion and Textual Recovery in the Undergraduate Classroom	
Earhart, Amy; Taylor, Toniesha	150
Potential Criticism in the Digital Humanities	
Edwards, Richard	152
Sequence, Tree and Graph at the Tip of Your Java Classes	
Eide, Øyvind	152
Exploratory Thematic Analysis for Historical Newspaper Archives	
Eisenstein, Jacob; Sun, Iris; Klein, Lauren F.	154
Literary Canon and Digital Bibliographies: The Case of the United States	
Ferrer, Carolina	156
μServices and The Riddle of Literary Quality	
Filarski, Gertjan; de Jong, Hayco; van Dalen-Oskam, Karina	156
From Markup to Analysis: Culture Claims and Code in the Digital Archive	
Flanders, Julia; Dillon, Elizabeth Maddock	157
Monolith: Materialised Bits, the Digital Rosetta Film	
Fornaro, Peter; Wassmer, Andreas; Rosenthaler, Lukas; Gschwind, Rudolf	159
Beyond Style: Literary Capitalism and the Publishing Industry	
Fuller, Simon; O'Sullivan, James	160
CAMPUS MEDIUS--Topography and Topology of a Media Experience	
Ganahl, Simon; Solomon, Rory; Brennan, Mallory; Daftary, Darius	163
The MAAYA Project: Multimedia Analysis and Access for Documentation and Decipherment of Maya Epigraphy	
Gatica-Perez, Daniel; Pallan, Carlos; Marchand-Maillet, Stephane; Odobez, Jean-Marc; Roman Rangel, Edgar; Grube, Nikolai	165
Automating the Search for Cross-language Text Reuse	
Gawley, James; Forstall, Christopher; Clark, Konnor	168

XML-Print. Typesetting arbitrary XML documents in high quality	169
Georgieff, Lukas; Küster, Marc Wilhelm; Selig, Thomas; Sievers, Martin	169
Let DH Be Sociological! [Short Paper]	
Goldstone, Andrew	171
Building a metrical ontology as a model to link digital poetic repertoires	
González-Blanco, Elena; Seláf, Levente; Del Rio Riande, María Gimena; Martínez Cantón, Clara Isabel; Martos Pérez, María Dolores	174
Exploring Usage of Digital Newspaper Archives through Web Log Analysis: A Case Study of Welsh Newspapers Online	
Gooding, Paul	175
DigCurV: curriculum framework for digital curation in the cultural heritage sector	
Gow, Ann; Molloy, Laura; Konstantelos, Leo	177
Digital approaches to understanding the geographies in literary and historical texts	
Gregory, Ian; Donaldson, Chris; Murrieta-Flores, Patricia; Rupp, C.J.; Baron, Alistair; Hardie, Andrew; Rayson, Paul	179
Topotime: Representing historical temporality	
Grossner, Karl; Meeks, Elijah	181
Does colour mean color?: Disambiguating word sense and ideology in British and American orthographic variants	
Grue, Dustin	183
Navigating the Storm: eMOP, Big DH Projects, and Agile Steering Standards	
Grumbach, Elizabeth; Christy, Matthew; Mandell, Laura; Neudecker, Clemens; Avil, Loretta; Samuelson, Todd; Antonacopoulos, Apostolos	184
Accessing, navigating, and engaging with high-resolution document image collections using Diva.js	
Hankinson, Andrew; Pugin, Laurent; Fujinaga, Ichiro	186
The Ancient Coins of Thrace: A Numismatic Web Portal	
Hanrahan, Elise	188
The Chimeria Platform: User Empowerment through Expressing Social Group Membership Phenomena	
Harrell, D. Fox; Kao, Dominic; Lim, Chong-U; Lipshin, Jason; Sutherland, Ainsley	189
Framework of an Advisory Message Board for Women Victims after Disasters	
Hashimoto, Takako; Shirota, Yukari	191
Quelle médiation numérique pour le patrimoine bâti ?	
Hennebert, Jérôme	193
Open content production in museums. A discourse and critical analysis of the museum in the digital age	
Hidalgo Urbaneja, María Isabel	195
CLARIN: Resources, Tools, and Services for Digital Humanities Research	
Hinrichs, Erhard; Krauwer, Steven	196
Trading Consequences: A Case Study of Combining Text Mining & Visualisation to Facilitate Document Exploration	
Hinrichs, Uta; Alex, Beatrice; Clifford, Jim; Quigley, Aaron	198
Tuning the Word Frequency List	
Hoover, David L.	200
Making Waves: Algorithmic Criticism Revisited	
Hoover, David L.	202
The Workspace for Collaborative Editing	
Houghton, Hugh; Sievers, Martin; Smith, Catherine	204
Enjambment and the Poetic Line: Towards a Computational Poetics	
Houston, Natalie	206
A glimpse of the change of worldview between 7th and 10th century China through two leishu	
Hsiang, Jieh; Chen, lihua; Chung, Chia-Hsuan	207
Building impact and value into the development of digital resources in the humanities: Rhyfel Byd 1914-1918 a'r profiad Cymreig / Welsh experience of World War One 1914-1918	
Hughes, Lorna; Roberts, Owain; McCann, Paul	209
Using digitized newspaper archives to investigate identity formation in long-term public discourse	
Huistra, Hieke; Pieters, Toine	210
Extracting Relationships from an Online Digital Archive about Post-War Queensland Architecture	
Hunter, Jane; Macarthur, John; Van der Plaat, Deborah; Gosseye, Janina; Muys, Andrae; Macnamara, Craig; Bannerman, Gavin	211
Student Collaborators in Digital Humanities Outreach and Advocacy: Strategies and Examples from the IDHMC at Texas A&M University	
Ives, Maura; Earhart, Amy; Grumbach, Elizabeth; Mandell, Laura	213
Using Social Network Analysis to Reveal Unseen Relationships in Medieval Scotland	
Jackson, Cornell Alexander	215
IMPACT : un dispositif de transcription et de commentaire de l'oral, pour l'enseignement et la recherche	
Jacquin, Jérôme; Gradoux, Xavier	216
Digital learning in an undergraduate context: promoting long term student-faculty (and community) collaboration in the Susquehanna Valley, PA	
Jakacki, Diane; Faull, Katherine	217
The Rowling Case: A Proposed Standard Analytic Protocol for Authorship Questions	
Juola, Patrick	218
5 Design Rules for Visualizing Text Variant Graphs	
Jänicke, Stefan; Geßner, Annette; Büchler, Marco; Scheuermann, Gerik	220
A Preparatory Analysis of Peer-Grading for a Digital Humanities MOOC	
Kaplan, Frédéric; Bornet, Cyril	222
WÆΓÑing: A Conceptual Parsing of ASCII Character Substitutions	
Katelnikoff, Joel	224
Problems in Encoding Documents of Early Modern Japanese	
Kawase, Akihiro; Ichimura, Taro; Ogiso, Toshinobu	225
History All Around Us: Towards Best Practices for Augmented Reality for Public History and Cultural Empowerment	
Kee, Kevin Bradley; Compeau, Timothy; Poitras, Eric	227

Aiding Modern Textual Scholarship using a Virtual Hinman Collator Kejriwal, Gaurav; Furuta, Richard; Olivieri, Ryan	228
Swiss Voice App: A smartphone application for crowdsourcing Swiss German dialect data Kolly, Marie-José; Leemann, Adrian; Dellwo, Volker; Goldman, Jean-Philippe; Hove, Ingrid; Almajai, Ibrahim	231
Beautiful lips and porcelain cheeks: extracting physical descriptions from recent Dutch fiction Koolen, Corina; Wubben, Sander; van Cranenburgh, Andreas	233
TheoPhilo. A prototype for a Thesaurus of Philosophy Lamarra, Antonio; Tardella, Michela	235
The social pleasure of the text: Applying digital humanities methods to reception studies Lang, Anouk	237
BFM Collection - Open-Source Digital Editions of Medieval French Texts Lavrentiev, Alexei	239
Modèles tridimensionnels pour la représentation de l'état des connaissances et propositions de visualisation pour l'analyse des corpus textuels. Leblanc, Jean-Marc; Pérès, Marie	240
Supporting "Distant Reading" for Web Archives Lin, Jimmy; Kraus, Kari; Punzalan, Ricardo L. Punzalan	241
Developing a Physical Interactive Space for Innovative Digital Humanities Exhibition Liu, Jyi-Shane; Liao, Wen-Hung	242
Mining the Cloud of Witness: Inferring the Prestige of Saints from Medieval Paintings Lombardi, Thomas	246
Detection of Poetic Content in Historic Newspapers through Image Analysis Lorang, Elizabeth M.; Soh, Leen-Kiat; Lunde, Joseph; Thomas, Grace	248
Visualizing Global News Losh, Elizabeth; Manovich, Lev	250
A Sense of Place: Mapping Fictional Landscapes in Literary Narratives Lynch, John; Kurtz, Wendy; Rocchio, Michael	253
Sentiment Analysis for the Humanities: the Case of Historical Texts Marchetti, Alessandro; Sprugnoli, Rachele; Tonelli, Sara	254
Circling around texts and language: towards 'pragmatic modelling' in Digital Humanities Marras, Cristina; Ciula, Arianna	257
STAK – Serendipitous Tool for Augmenting Knowledge: Bridging Gaps between Digital and Physical Resources Martin, Kim; Greenspan, Brian; Quan-Haase, Anabel	259
Designing the next big thing: Randomness versus serendipity in DH tools Martin, Kim; Quan-Haase, Anabel	261
Small-Scale Big Data: Experimental Literature and Distributed Computing Mauro, Aaron	263
Pushing Back the Boundary of Interpretation: Concept, Practice and Relevance of a Digital Heuristic Meister, Jan Christoph; Jacke, Janina	264
Visualizing Computational, Transversal Narratives from the World Trade Towers Miller, Ben; Shrestha, Ayush; Olive, Jennifer	266
Clustering Search to Navigate A Case Study of the Canadian World Wide Web as a Historical Resource Milligan, Ian	268
The Telltale Hat: LDA and Classification Problems in a Large Folklore Corpus Mimno, David; Broadwell, Peter M.; Tangherlini, Timothy R.	271
Seeing the Trees & Understanding the Forest Montague, John Joseph; Rockwell, Geoffrey; Ruecker, Stan; Sinclair, Stéfan; Brown, Susan; Chartier, Ryan; Frizzera, Luciano; Simpson, John	272
Making Digital Humanities Work Munoz, Trevor; Giuliano, Jennifer	274
Tracking Semantic Drift in Ancient Languages: The Bible as Exemplar and Test Case Munson, Matthew	275
Bridging the Local and the Global in DH: A Case Study in Japan Nagasaki, Kiyonori; Muller, A. Charles; Tomabechi, Toru; Shimoda, Masahiro	277
Active Authentication through Psychometrics Noecker Jr, John	278
Encoding Metaknowledge for Historical Databases Nuessli, Marc-Antoine; Kaplan, Frédéric	280
Two Irish Birds: A Stylometric Analysis of James Joyce and Flann O'Brien O'Sullivan, James; Bazarnik, Katarzyna; Eder, Maciej; Rybicki, Jan	281
Modeling Linguistic Research Data for a Repository for Historical Corpora Odebrecht, Carolin	284
A hipersensibilidade do Território – viver entre terra e nuvens Oliveira, Lídia; Baldi, Vania	285
MapaHD: Exploring Spanish and Portuguese Speaking DH Communities Ortega, Érika; Gutiérrez, Silvia	288
Geoweb 2.0 and Design Empowerment: A Critical Evaluation of Eleven Cases Pak, Burak; Verbeke, Johan	290
A vocabulary of the aesthetic experience for modern dance archives Paquette-Bigras, Ève; Forest, Dominic	292
Mixing contributions, collaborations and co-creation: participatory archaeology through crowd-sourcing Pett, Daniel Edward John; Bonacchi, Chiara; Bevan, Andy	294
Treasure Challenge: an archaeological video conferencing journey Pett, Daniel Edward John; Kelland, Katharine Louise	295
On Metaphor in Text Visualization Prototypes Peña, Ernesto; Brown, Monica; Dobson, Teresa	296

Modelling digital editing: of texts, documents and works	
Pierazzo, Elena; Noël, Geoffroy	298
Cultural text mining: using text mining to map the emergence of transnational reference cultures in public media repositories	
Pieters, Toine; Verheul, Jaap	299
Aplicación del análisis dinámico de redes científicas al estudio de la evolución de la investigación española relacionada con el descriptor "historia del arte" durante 1976-2012, según ISOC.	
Pino-Díaz, José; Cruces-Rodríguez, Antonio; Rodríguez-Ortega, Nuria; Bailón-Moreno, Rafael	301
Incommensurability? Authorship, Style, and the Need for Theory	
Plasek, Aaron	303
Starting the Conversation: Literary Studies, Algorithmic Opacity, and Computer-Assisted Literary Insight	
Plasek, Aaron; Hoover, David L.	305
Who is we? The social media project: Día de las humanidades digitales/Dia das humanidades digitais	
Priani, Ernesto; Spence, Paul; Galina Russell, Isabel; González-Blanco, Elena; Paixão de Sousa, Maria Clara; Alves, Daniel; Barrón, José Francisco; Godinez, Marco Antonio; Guzmán, Ana María	306
Reconstruction and Display of a Nineteenth Century Landscape Model	
Priestnall, Gary; Katharina, Lorenz; Mike, Heffernan; Joe, Bailey; Craig, Goodere; Robyn, Sullivan	308
Constructing Scientific Archives that Support Humanistic Research	
Prom, Christopher	309
Digital Linguistic Archive of the Dutch East India Company (VOC): Modeling a community-sourcing platform for historical linguistic research	
Pyltowany, Anna	312
On automatically disambiguating end-of-line hyphenated words in French texts	
Pöckelmann, Marcus; Ritter, Julia; Gießler, André	313
Fractures and Cohesion: Using Systemic Functional Linguistics to Detect and Analyse Hate Speech in an Online Environment	
Quinn, Deirdre; Maycock, Keith; Keating, John	315
Framework for Quantitative Analysis of Scripts	
Rajan, Vinodh	317
Macro-Etymological Textual Analysis	
Reeve, Jonathan Pearce	319
Crowdsourcing Performing Arts History with NYPL's ENSEMBLE	
Reside, Doug	320
Two new tools for multimodal editions	
Reside, Doug	321
Using Computer Vision to Improve Image Metadata	
Reside, Doug	323
Introducing digital humanities through the analysis of cultural productions	
Reyes-Garcia, Everardo	324
The Story of Stopwords: Topic Modeling an Ekphrastic Tradition	
Rhody, Lisa	326
Play as Process and Product: On Making Serendip-o-matic	
Ridge, Mia; Croxall, Brian; Papaelias, Amy; Kleinman, Scott	327
Harddrive Philology: Analysing the Writing Process on Thomas Kling's Archived Laptops	
Ries, Thorsten	329
A Network Analysis Approach of the Venetian Incanto System	
Rochat, Yannick; Fournier, Melanie; Mazzei, Andrea; Kaplan, Frédéric	330
Canon, value and artistic culture: critical inquiry about the new processes of assigning value in the digital realm	
Rodríguez-Ortega, Nuria	331
Discourses and Disciplines in the Enlightenment: Topic Modeling the French Encyclopédie	
Roe, Glenn; Gladstone, Clovis; Morrissey, Robert	334
National Data Curation and Service Center for Digital Research Data in the Humanities	
Rosenthaler, Lukas; Fornaro, Peter; Clivaz, Claire	335
Mixed data, mixed audience: building a flexible platform for the Visionary Cross project	
Rosselli Del Turco, Roberto	337
Simulating the Cultural Evolution of Literary Genres	
Sack, Graham Alexander; Wu, Daniel; Zusman, Benji	339
Computational Models of Narrative: Using Artificial Intelligence to Operationalize Russian Formalist and French Structuralist Theories	
Sack, Graham Alexander; Finlayson, Mark; Gervas, Pablo	340
An Integrated Approach to the Procedural Modeling of Ancient Cities and Buildings	
Saldana, Marie	342
Digital Humanities Empowering through Arts and Music. Tunisian Representations of Europe through music and video clips	
Salzbrunn, Monika; Mastrangelo, Simon	343
Digital humanities in Estonia: digital divide or linguistic isolation?	
Sarv, Mari; Kulasalu, Kaisa	345
"How To Do (Digital) History" and Undergraduate Digital Humanities	
Schell, Justin; Gabaccia, Donna	346
Intellectual Property Rights vs. Freedom of Research: Tripping stones in international IPR law	
Scholger, Walter	347
Revisionism as Outreach: The Letters of 1916 Project	
Schreibman, Susan	348
Digitizing the Dead and Dismembered: DH Technologies for the Study of Coptic Texts	
Schroeder, Caroline T.; Zeldes, Amir	349
Progress Through Regression. Modeling Style across Genre in French Classical Theater	
Schöch, Christof; Riddell, Allen	350
Hartmut Skerbisch – Envisioning association processes of a conceptual artist	
Semlak, Martina	352
The potential of open computer-mediated communication channels to facilitate collaboration in geographically distributed collaborations	
Siemens, Lynne	353

Pelagios 3: Towards the semi-automatic annotation of toponyms in early geospatial documents	355
Simon, Rainer; Barker, Elton T. E.; de Soto, Pau; Isaksen, Leif	355
Towards an Archaeology of Text Analysis Tools	356
Sinclair, Stéfan; Rockwell, Geoffrey	356
Advocating for a Digital Humanities Curriculum: Design and Implementation	358
Smith, David	358
Transcriptional implicature: a contribution to markup semantics	360
Sperberg-McQueen, Michael; Marcoux, Yves; Huitfeldt, Claus	360
Arch-V: A Platform for Image-Based Search and Retrieval of Digital Archives	362
Stahmer, Carl	362
Digital Pedagogy is about Breaking Stuff	362
Stommel, Jesse	362
Creating a Digital Tombstone Archive: From Fieldwork to Theory Formation	364
Streiter, Oliver; Goudin, Yoann	364
Future Development of a System for Annotation and Linkage of Sources in Arts and Humanities	366
Subotic, Ivan; Kilchenmann, André; Schweizer, Tobias; Rosenthaler, Lukas	366
A Large Database Approach to Cultural History	368
Sullivan, Brenton	368
Digitizing Women's Literary History: The Possibility Of Collaborative Empowerment?	368
Suzan, van Dijk; Dekker, Ronald; Partzsch, Henriette; Prats Lopez, Montserrat; Sanz, Amelia; Filarski, Gertjan	368
Analysis of perspectives in contemporary Japanese novels using computational stylistic methods	370
Suzuki, Takafumi; Yamashita, Natsumi	370
Integrating Score and Sound: "Augmented Notes" and the Advent of Interdisciplinary Publishing Frameworks	372
Swafford, Joanna	372
Realizing the democratic potential of online sources in the classroom	374
Sweeny, Robert C.H.; Burton, Valerie C.	374
Stylometry of Collaborations: Dickens, Collins and their collaborative writings	375
Tabata, Tomoji	375
Towards a Semantic Network of Dante's Works and Their Contextual Knowledge	378
Tavoni, Mirko; Andriani, Paola; Bartalesi, Valentina; Locuratolo, Elvira; Meghini, Carlo; Versienti, Loredana	378
A "Deeply Annotated" Bibliography of Local Social Histories of Early Modern Europe	379
Theibault, John Christopher	379
What remains to be done – Exposing invisible collections in the other 6500 languages and why it is a DH enterprise.	380
Thieberger, Nick	380
Photogrammar: Organizing Visual Culture through Geography, Text Mining, and Statistical Analysis	382
Tilton, Lauren; Leonard, Peter; Arnold, Taylor	382
A novel approach for a reusable federation of research data within the arts and humanities	382
Tobias, Gradl; Henrich, Andreas	382
Problems in Modeling Transactions	385
Tomasek, Kathryn; Bauman, Syd	385
When Kidnapping is but One Risk: Digital Studies Challenge Scholarly and Regional Cultures	386
Toth, Michael; Emery, R. Douglas	386
"Needless To Say": Articulating Digital Publishing Practices as Strategies of Cultural Empowerment	388
Tullos, Allen E.	388
Relating texts to 3D-information: A generic software environment for Spatial Humanities	389
Unold, Martin; Lange, Felix	389
Dynamic Visualizations In Enriched Publications Of Seventeenth Century Science	391
Van den Heuvel, Charles; Cocquyt, Tiemen; Hoogerwerf, Maarten; Nagel, Dylan; Thijssen, Michiel	391
Process Data for Digital Scholarly Editions	393
Vasold, Gunter	393
The opportunistic librarian: A Leuven confession	395
Verbeke, Demmy	395
Kinomatics: big cultural data and the study of cinema	396
Verhoeven, Deb; Coate, Bronwyn; Arrowsmith, Colin; Davidson, Alwyn	396
Less explored multilingual issues in the automatic processing of historical texts – a case study	397
Vertan, Cristina	397
Modelling digital edition of medieval and early modern accounting documents	398
Vogeler, Georg	398
Archaeology in social media: users, content and communication on Facebook	400
Vosyliute, Ingrida	400
Digital Humanists Are Motivated Annotators	401
Walkowski, Niels-Oliver; Barker, Elton T. E.	401
The dog that didn't bark: A longitudinal study of reading behaviour in physical and digital environments	402
Warwick, Claire; Mahony, Simon; Rayner, Samantha; Team, The INKE	402
Ideas, Events and Actions: The Digital Humanity Study of the Concept Formation in Modern China	404
Wen-huei, Cheng; Jui-sung, Yang; Wei-Yun, Chiu; Chao-lin, Liu; Guan-tao, Jin; Qing-feng , Liu	404
Lacuna Stories: Building an Annotation Platform for Historical Thinking	405
Widner, Michael; Johnsrud, Brian	405
Building the social graph of the History of European Integration: A pipeline for the Integration of Human and Machine Computation	407
Wieneke, Lars; Sillaume, Ghislain; Düring, Marten; Pasini, Chiara; Fraternali, Piero; Tagliasacchi, Marco; Melenhorst, Marc; Novak, Jasminko; Micheel, Isabel; Harloff, Erik; Garcia Moron, Javier; Lallemand, Carine; Vincenzo, Croce; Lazzaro, Marilena; Nucci, Francesco	407
Kanripo and Mandoku: Tools for git-based distributed repositories for premodern Chinese texts	408
Wittern, Christian	408
A Morphological Analysis of Classical Chinese Texts	409
Yasuoka, Koichi; Yamazaki, Naoki; Wittern, Christian; Nikaido, Yoshihiro; Morioka, Tomohiko	409

Collaboratively maximizing inter-ontology agreement for controversial domains: A case study of Jewish cultural heritage Zhitomirsky-Geffet, Maayan; Erez, Eden Shalom	411
The Changing Canon of Beauty: Facial Attractiveness in the Representation of Human Faces in World Painting de la Rosa, Javier; Caldas, Natalia; Dutta, Nandita; Suárez, Juan Luis	413
ClipNotes: Digital Annotation and DataMining for Film & Television Analysis deWaard, Andrew	415
Distributed "Forms of Attention": eMOP and the Cobre Tool duPlessis, Anton Raymund; Mandell, Laura; Creel, James; Maslov, Alexey	416
The Problem of Time and Space: The Difficulties in Visualising Spatiotemporal Change in Historical Data Ó Murchú, Tomás; Lawless, Séamus	419

Posters

Mapping French Press to the Digital Age Abi Haidar, Alaa; Ganascia, Jean-Gabriel	422
A Digital Metaphor Map for English Anderson, Wendy; Aitken, Brian; Hamilton, Rachael	424
Quantifying "The Thing Not Named": A Computational Analysis of Willa Cather's "Unfurnished" Writing Style(s) Ankenbrand, Rebecca; Bernardini, Caterina; Brotnov Eckstrom, Mikal; Kinnaman, Alex; Tedrow, Kimberly Ann	425
Le labo junior « Nhumérisme » (ENS Lyon), observateur et acteur du « cultural empowerment » français Armand , Cécile	426
Light, Liturgy, and Art at the Monastery of Saint John in Müstair, Switzerland: A Software Demonstration Ataoguz, Kirsten	427
Interoperable Infrastructures for Digital Research: a proposed pathway for enabling transformation Baker, James; Farquhar, Adam	427
Neue Möglichkeiten der Arbeit mit strukturierten Sprachressourcen in den Digital Humanities mithilfe von Data-Mining Bartz, Thomas; Beißwenger, Michael; Pöltz, Christian; Radke, Nadja; Storrer, Angelika	428
Mapping Colonial Americas Publishing Project Bauer, Jean; Egan, James	431
Data Curation Nightmare: Migrating VM/CMS to GNU/Linux in 2 weeks Bauman, Syd; DiCamillo, Peter	431
The Open Philology Project at the University of Leipzig Baumgardt, Frederik; Berti, Monica; Celano, Giuseppe; Crane, Gregory R.; Dee, Stella; Foradi, Maryam; Franzini, Emily; Franzini, Greta; Stoyanova, Simona	432
What's in a Discipline? Research Practices, Use of Tools and Content in the Humanities and Social Sciences - The web-based questionnaires of EHRI and Europeana Cloud. Benardou, Agiatis; Papaki, Eliza; Chatzidiakou, Nephelie	433
The CENDARI Project: A user-centered 'enquiry environment' for modern and medieval historians Benes, Jakub; O'Connor, Alex; Dimara, Evanthis	434
Mountains of Text. Analyzing Alpine Literature from the AAC Biber , Hanno	436
Exploring Qualitative Data for Secondary Analysis: Challenges, Methods, and Technologies Bischoff, Kerstin; Niederée, Claudia; Tran, Nam Khanh; Zerr, Sergej; Birke, Peter; Brückweh, Kerstin; Wiede, Wiebke	436
SNAP:DRGN - Standards for Networking Ancient Prosopographies: Data and Relations in Greco-roman Names Bodard, Gabriel; Depauw, Mark; Rahtz, Sebastian	439
Probing Digital Scholarly Curation through the Dynamic Table of Contexts Brown, Susan; Adelaar, Nadine; Dobson, Teresa; Knechtel, Ruth; MacDonald, Andrew; Nelson, Brent; Peña, Ernesto; Radzikowska, Milena; Roeder, Geoff G.; Ruecker, Stan; Sinclair, Stéfan; Windsor, Jennifer; INKE Research Group,	439
The CWRC-Writer Bridge: from Coder to Writer, XML to RDF, DH to Mainstream Brown, Susan; Brundin, Michael; Chartrand, James; Knechtel, Ruth; MacDonald, Andrew; Rockwell, Geoffrey; Sellmer, Megan	441
Transcribo: A Graphical Editor for Transcribing and Annotating Textual Witnesses. Preparing a Historical-Critical Edition of Arthur Schnitzler's Works. Buedenbender, Stefan; Friedrich, Vivien; Burch, Thomas; Fink, Kristina; Wolfgang, Lukas; Kathrin, Nühlen; Frank, Queens; Joshgun, Sirajzade	443
Digitalizing the Matsu Festival Celebration: The Study and Application of Value-Added Creative Methods to Taiwan Folk Culture and Art Chen, Chun-Wen; Hsu, Su-Chu; Day, Jia-Ming; Lin, Cheng-Wei	444
Arabic and Greek New Testament manuscripts: Identities and Digital cultures Clivaz, Claire; Schulthess, Sara; Bouvier, David; Teule, Herman	445
Empowering Play, Experimenting with Poems: Disciplinary Values and Visualization Development Coles, Katharine; Meyer, Miriah; Lein, Julie Gonnering; McCurdy, Nina	446
CatCor: Correspondence of Catherine the Great Cummings, James; Rubin-Detlev, Kelsey; Kahn, Andrew	448
Empowering The Matsu(Goddess) Festival Celebration: From Static Woodcut Print to Animated Art Day, Jia-Ming; Hsu, Su-Chu	449
Visualizing theatrical heritage: Computer modelling as a tool for researching the theatre history of the Low Countries De Paepe, Timothy	450
Orthography and Biblical Criticism Dershowitz, Idan; Dershowitz, Nachum; Hasid, Tomer; Ta-Shma, Amnon	451
L'Innommable / The Unnamable: The Second Module of the Samuel Beckett Digital Manuscript Project's Hybrid Genetic Edition. Dillen, Wout	453
Spreading DiRT: extending the Digital Research Tools directory Dombrowski, Quinn; Gold, Matthew	453

An easy tool for creating digital scholarly editions	454
Dumont, Stefan; Fechner, Martin	454
Créer un centre de recherche interuniversitaire sur les humanités numériques au Québec : Défis et succès	455
Eberle-Sinatra, Michael; Sinclair, Stéfan; Dyens, Olliver; Vitali Rosati, Marcello	455
Stylometry, network analysis, and Latin literature	456
Eder, Maciej	456
Network Analysis for the History of Religions – The SeNeReKo project	457
Elwert, Frederik; Wortmann, Sven; Hofmann, Beate; Knauth, Jürgen	457
Taking manuscripts apart, and putting them together	458
Emery, R. Douglas; Porter, Dot; Campagnolo, Alberto	458
Visualizing Homelessness	459
Engel, Maureen; Zwicker, Heather; Frizzera, Luciano; Pedraça, Samia; Regattieri, Lorena; Schoenberger, Zachary; Windsor, Jennifer	459
Digital Actors' Parts: An Interactive Tool for Learning Shakespeare's Plays	460
Estill, Laura; Meneses, Luis	460
Combining Macro- and Microanalysis for Exploring the Construal of Scientific Disciplinarity	461
Fankhauser, Peter; Kernes, Hannah; Teich, Elke	461
Reading Again: Annotating, Editing, and Writing in the Browser. Pedagogy, Design, and Development of Annotation Studio	463
Fendt, Kurt; Kelley, Wyn; Folsom, Jamie	463
Rethinking HathiTrust Metadata to Support Workset Creation for Scholarly Analysis	465
Fenlon, Katrina; Cole, Timothy; Han, Myung-Ja; Willis, Craig; Fallaw, Colleen	465
Enhancing Scholarly Communication and Communities with the PressForward Plugin	466
Fragaszy Troyano, Joan; Rhody, Lisa; Coble, Zach; Shirazi, Roxanne; Potvin, Sarah; Pinto, Caro	466
Check! – An online tool for the recognition and evaluation of DH work	467
Galina Russell, Isabel; Priani Saisó, Ernesto	467
The SyMoGIH project : Sharing and publishing historical and geographical data in a standard, open and interoperable way	468
Gedzelman, Séverine, Sonia; Beretta, Francesco; Ferhod, Djamel; Boschetto, Sylvain; Butez, Claire-Charlotte; Vernus, Pierre; Hours, Bernard	468
User-friendly lemmatization and morphological annotation of Early New High German manuscripts	469
Gießler, André; Ritter, Jörg; Molitor, Paul; Andert, Martin; Kösser, Sylvia; Leipold, Aletta	469
The MiCLUES system: Dynamic, rich contextual support for museum visits	471
Gold, Nicolas E.; Rossi Rognoni, Gabriele	471
How a story is performed: traditional storytelling in the hands of computing	472
Gomes, Mariana	472
The Early Modern OCR Project (eMOP): Fostering Access to Early Modern Cultural Materials	473
Grumbach, Elizabeth; Mandell, Laura; Christy, Matthew	473
The Annotated Star: A Collaborative Digital Edition of Rosenzweig's Star of Redemption	474
Handelman, Matthew; Wygoda, Ynon; Rojansky, Shay; Rusinek, Sinai	474
The SMART-GS Project: An Approach to Image-based Digital Humanities	475
Hashimoto, Yuta; Aihara, Kenro; Hayashi, Susumu; Kukita, Minao; Ohura, Makoto	475
Open-Access Cultural Heritage Resources and Native American Stakeholders: A Case Study from Chaco Canyon, New Mexico	476
Heitman, Carrie C.	476
Collaborative Scholarly Building with the Early Caribbean Digital Archive	477
Hopwood, Elizabeth; Doyle, Benjamin	477
The Development of The Dickens Lexicon Digital and its Practical Use for the Study of Late Modern English	477
Hori, Masahiro; Imabayashi, Osamu; Tabata, Tomoji; Koguchi, Keisuke; Nishio, Miyuki; Nagasaki, Kiyonori	477
What we make of Code: The Role of Programming in the Digital Humanities	478
Jakacki, Diane; O'Sullivan, James	478
Measuring the style of chick lit and literature	480
Jautze, Kim Johanna	480
All databases are created equal: building profiles for database standards and interoperability in the Humanities	482
Johnson, Ian R.	482
Local voices, worldwide conversations: ethnographic methodologies as a route to understanding meaning and value of niche local digital cultural heritage resources.	483
Johnston, Penny	483
Data Criticism: General Framework for the Quantitative Interpretation of Non-Textual Sources	484
Kitamoto, Asanobu; Nishimura, Yoko	484
Shedding Light on Dickens' Style Through Representativeness & Distinctiveness	485
Klaussner, Carmen; Nerbonne, John; Çöltekin, Çağrı	485
Finding Inexact Quotations Within a Tibetan Buddhist Corpus	486
Klein, Benjamin Eliot; Dershowitz, Nachum; Lior, Wolf; Almogi, Orna; Wangchuk, Dorji	486
Supporting cross-media analyses by automatically linking multiple collections	488
Kleppe, Martijn; Kemman , Max	488
LAUDATIO-Repository: Accessing a heterogeneous field of linguistic corpora with the help of an open access repository	489
Krause, Thomas; Lüdeling, Anke; Odebrecht, Carolin; Romary, Laurent; Schirmbacher, Peter; Zielke, Dennis	489
Detecting Linguistic Signal in Cather's Early Journalism: Polishing the Bibliography	490
Kumari, Ashanka; Lawton, Courtney; McCue, Carmen; Moreno, Jose Luis; Thomas, Grace	490
Annotating texts with ontologies, from geography to persons and events	492
Lana, Maurizio; Ciotti, Fabio; Magro, Diego; Peroni, Silvio; Tomasi, Francesca; Vitali, Fabio	492
Enduring Traces: Exploring correspondence from the archives of Canadian modernism using digital tools and methods	494
Lang, Anouk	494
Cultivating the Public Philosophy Journal	495
Long, Christopher; Fisher, Mark; Rehberger, Dean	495
Library Science and Textual Transmission in the Online Age: A Fluid Text Model and Proposed Documenting Infrastructure	497
MacCall, Steven L.	497
The Inherited Self: Reappraising Literary Cultural Heritage through Digital Methods	497
Malm, Mats Ulrik; Bergemar, Jenny; Kokkinakis, Dimitrios; Leonard, Peter	497

Liberate the Text! TypeWright, Cobre, and MapThePage	
Mandell, Laura C.; Heil, Jacob; Duguid, Timothy; Grumbach, Elizabeth; Christy, Matthew	498
Building the Princeton Prosody Archive	
Martin, Meredith; Wythoff, Grant; Wilson, Meagan; Brown, Travis	499
TEI Customization for encoding paratexts in spanish printed books (XV-XVIII)	
Martos Pérez, María Dolores; Baranda Leturio, Nieves; Marín Pina, Mª Carmen	500
Edition Visualization Technology: a simple tool to publish digital editions and digital facsimiles	
Masotti, Raffaele; Kenny, Julia; Di Pietro, Chiara	501
Our Marathon: The Boston Bombing Digital Archive	
McGrath, Jim; Peaker, Alicia	502
Speaking in Code	
Nowviskie, Bethany; Rochester, Eric; Graham, Wayne; Boggs, Jeremy; McClure, David; Bailey, Scott	502
La Hiperedição Dos Panfletos De Eulálio Motta: Edición Filológica Y Cultura Digital, Retos De Un Nuevo Tiempo	
Nunes Barreiros, Patrício	503
The Digitization of Hmong Sacred Texts	
Ogden, Mitchell Paul	506
Digital Humanities as Vocation: Possibilities for Undergraduate Education	
Ogden, Mitchell Paul	506
Modeling Melville's Reading: Editing Marginalia in TEI, Topic Modeling Reading and Influence	
Ohge, Christopher M.	507
Large-scale text analysis through the HathiTrust Research Center	
Organisciak, Peter; Bhattacharyya, Sayan; Auvin, Loretta; Plale, Beth; Downie, J. Stephen	507
Critical editing with TXSTEP	
Ott, Wilhelm; Ott, Tobias	509
Visualization As a Bridge to Close Reading: The Audience in <i>The Castle of Perseverance</i>	
Peterson, Noah	513
How we work: a critical approach to program development to serve library/dh partnerships	
Potvin, Sarah; Herbert, Bruce; Earhart, Amy	514
Seeing Dialogue: Network Visualization of Dramatic Texts	
Powell, Daniel	516
Discovering Old Maps Online and Transforming Them Into Digital Humanities Resources	
Pridal, Petr	517
Geographies of Access: Mapping the Online Attention to Digital Humanities Articles in Academic Journals	
Priego, Ernesto; Havemann, Leo; Atenas, Javiera	517
Big Data and the Literary Archive: Topic Modeling the Watson-McLuhan Correspondence	
Quamen, Harvey; Hjartarson, Paul	518
The Digital Alchemist: A Mixed Reality Exploration of Jonson's Alchemist as Site-Specific Theatre	
Quinsland, Kirk; Rouse, Rebecca	519
How to make games more GLAM-orous: developing game prototypes for the museum and cultural heritage sector in India	
Ray Murray, Padmini	521
Enhancing Access to Online Oral History: Oral history in the Digital Age (OHDA) and Oral History Metadata Synchronizer (OHMS)	
Rehberger, Dean; Boyd, Douglas	522
Marked E-Books and Kindle's popular highlight culture	
Rowberry, Simon	522
Sustainability?! Four Paradigms for Humanities Data Centers	
Sahale, Patrick; Kronenwett, Simone; Blumtritt, Jonathan	523
Euterpe's Hidden Song: Patterns in Elegy	
Scheirer, Walter; Forstall, Christopher	524
DARIAH-DE – Digital Infrastructure for the Arts and Humanities	
Schmunk, Stefan; Smith, Kathleen; Blümm, Mirjam	526
The Text portal: An online resource providing medieval literature for students and their teachers	
Schneider, Gerlinde; Schwinghammer, Ylva	527
Digital Humanities Data Curation Institutes: Challenges and Preliminary Findings	
Senseney, Megan; Muñoz, Trevor; Flanders, Julia; Fenlon, Ali	528
Adams Family Legacy: Visualizing the World of an American Presidential Family	
Sikes, Sara; Christian-Lamb, Caitlin	529
Empowering Student Digital Scholarship: CLASS Program as a model for digital humanities scholarship in the Liberal Arts	
Simons, Janet Thomas; Nieves, Angel David; Grimaldi, Kerri	529
Reading Between the Lines: Image-to-Segment Relationship Development and Analysis	
Smith, Dustin; Karadkar, Unmil; Galloway, Pat; Davis, King	533
Taking a Global Perspective on the Skills and Competencies Important to Digital Scholarship	
Spiro, Lisa; Cawthorne, Jon; Lewis, Vivian; Wang, Xuemao	534
Converting Medieval Documents into a Searchable Database	
Sporleder, Caroline; Fertmann, Susanne; Krones, Tim; Kolatzek, Robert; Teufel, Isolde	535
Interdisziplinarität modellieren – Über die Modellierung einer Ontologie wissenschaftlicher Prozesse für den Exzellenzcluster Bild Wissen Gestaltung	
Stein , Christian	537
Cirilo Client: An application for data curation and content preservation	
Steiner, Elisabeth	539
The DigiPal Framework for Script and Image	
Stokes, Peter A.; Brookes, Stewart; Noël, Geoffroy; Buomprisco, Giancarlo; Marques de Matos, Debora; Watson, Matilda	539
Digital multi-text editions from scratch to electronic performance. Transcription and collation routines transformed in a flexible database system	
Stoltz, Michael	541
Medical Humanities. Projet de musée digital	
Suciú, Radu; Wenger, Alexandre; Bolli, Laurent	542

Visualization of Historical Knowledge Structures: An Analysis of the Bibliography of Philosophy Sula, Chris Alen; Dean, Will	543
TextGrid: Creating, archiving, publishing and exploring digital editions and other humanistic research data via a Virtual Research Environment Söring, Sibylle; Veentjer, Ubbo; Funk, Stefan	544
The Arabic Papyrology Database Thomann, Johannes	545
"Crowdsourcing Annotation and the 'Social Edition': Ossian Online." Tonra, Justin; Barr, Rebecca	546
A Quantitative Analysis for the Authorship of Saikaku's Posthumous Works in the Seventeenth Century Uesaka, Ayaka; Murakami, Uesaka	547
The Proportional Sizes of Genres in Eighteenth- and Nineteenth-Century English-Language Books Underwood, Ted	549
Kiln: XML Publishing Framework Vieira, Miguel; Norrish, Jamie	550
An ontology for 3D visualisation of cultural heritage Vitale, Valeria	551
Linked Open Data Technologies and Emblematica Online II Wade, Mara R; Cole, Timothy; Han, Myung-Ja	552
Visualization, Interactivity and Contextualization as Digital Cultural Empowerment: Ancient Egyptian Architectural Terminology OnlinePaper created on 2013-11-01, 08:52 Wendrich, Willeke	553
HistoGlobe - Visualising History Westermeier , Carola	553
Semantic Blumenbach: Exploration of text-object relationship with semantic web technologies in the history of science Wettlaufer, Jörg	554
"Tout ce qui n'est point vers, est prose" : Raymond Queneau's Matrix Analysis of Language, Syntactic Stylometry, and Exploratory Programming Wolff, Mark	555
A Text Encoding Support System for Pre-modern Japanese Historical Materials Yamada, Taizo; Inoue, Satoshi	558
Taco: A Metadata System for Hierarchical Structured Data Collections Zastrow, Thomas; Gross, Karin	559

List of Reviewers

Alexander Marc	Dawson John	Keramidas Kimon
Alsop Peter Roger	Deegan Marilyn	Kermanidis Katia Lida
Alvarado Rafael	DiNunzio Joseph	Kleppe Martijn
Anderson Deborah	Dunn Stuart	Knox Douglas
Anderson Wendy	Dunning Alastair	Koh Adeline
Andreev Vadim Sergeevich	Earhart Amy	Körner Fabian
Andrews Tara Lee	Eckart Thomas	Kretzschmar William
Antonijevic Smiljana	Eckert Kai	Krot Michael Adam
Appleford Simon James	Eder Maciej	Lana Maurizio
Armand Cécile	Edmond Jennifer C	Lang Anouk
Arneil Stewart	Eide Øyvind	Lavagnino John
Arthur Paul	Engel Deena	Lavrentiev Alexei
Ashton Andrew Thomas	Ensor Jason	Lawless Séamus
Audenaert Michael Neal	Escandell-Montiel Daniel	Lawrence Katharine Faith
Baillot Anne	Esteva Maria	Levallois Clement
Bamman David	Fendt Kurt E	Litta Modignani Picozzi Eleonora
Barney Brett	Ferschke Oliver	Lombardini Dianella
Bartsch Sabine	Finn Edward	Lorang Elizabeth M
Baudoin Patsy	Fischer Franz	Losh Elizabeth
Bauer Jean Ann	Fitzpatrick Kathleen	Lüngen Harald
Bauman Syd	Flanders Julia	Lyman Eugene W
Baumann Ryan Frederick	Forest Dominic	Maeda Akira
Beavan David	Fournier Mélanie	Mahony Simon
Bennis Hans	Fraistat Neil R.	Makinen Martti
Berra Aurélien	France Fenella Grace	Maly Kurt
Biber Hanno	French Amanda	Mancera Rueda Ana
Bingenheimer Marcus	French Scot	Mandell Laura C
Blanke Tobias	Furner Jonathan	Martin Kim
Boggs Jeremy	Furuta Richard	Martin Worthy N.
Boot Peter	Galina Russell Isabel	Martín Arista Javier
Bordalejo Barbara	Gallet-Blanchard Liliane	McCarty Willard
Borgman Christine L.	Garfinkel Susan	McDaniel Rudy
Borin Lars	Gärtner Kurt	McDonald Jarom Lyle
Bornet Cyril	Gartner Richard	McTesterson Testy
Boschetti Federico	Gibbs Fred	Meschini Federico
Bosse Arno	Gil Alexander	Miles Adrian
Bouchard Matthew	Girard Paul	Mimno David
Boyd Jason Alexander	Gist D. Chris	Miyake Maki
Bradley John	Goetz Sharon K.	Monteiro Vieira Jose Miguel
BRISAC Anne-Laure	Gold Matthew K.	Morrison Aimée
Brown James	González-Blanco Elena	Moulin Claudine
Brown Susan	Gooding Paul Matthew	Moulthrop Stuart
Brown Travis Robert	Gow Ann	Mueller Martin
Büchler Marco	Gradmann Stefan	Mühlberger Günter
Burr Elisabeth	Graham Wayne	Muñoz Trevor
Burrows Toby Nicolas	Green Harriett Elizabeth	Munson Matthew Aaron
Buzzetti Dino	Gregory Ian	Muralidharan Aditi
CANTEAUT Olivier	Guertin Carolyn	Murphy Orla
Caton Paul	Guiliano Jennifer Elizabeth	Mylonas Elli
Cayless Hugh	Hankins Gabriel Anderson	Nagasaki Kiyonori
Cheesman Tom	Harris Katherine D.	Nelson Brent
Chen Shih-Pei	Hawkins Kevin Scott	Nerbonne John
Chue Hong Neil	Heath Sebastian	Nikiforova Larisa
Ciula Arianna	Heiden Serge	Norrish Jamie
Clavaud Florence	Henny Ulrike Edith Gerda	Nowviskie Bethany
Clavert Frédéric	Hettel Jacqueline	Nyhan Julianne
Clement Tanya	Heuvel Charles van den	O'Connor Alexander
Clivaz Claire	Heyer Gerhard	O'Donnell Daniel Paul
Conway Paul	Hirsch Brett	O'Sullivan James Christopher
Cordell Ryan	Ho Hou leong	OHYA Kazushi
Cowan William	Holmes Martin	Olsen Mark
Craig Hugh	Hoover David L.	Ore Christian-Emil
Crawford Tim	Howell Sonia Marian	Ore Espen S.
Crompton Constance	Hsiang Jieh	Organisciak Peter
Croxall Brian	Hunter Jane	Ortega Elika
Cummings James C.	Hunyadi László	Perdue Susan Holbrook
Cunningham Richard	Isaksen Leif	Perez Isasi Santiago
Dacos Marín	Ivanovs Aleksandrs	Pierazzo Elena
Dahlstrom Mats	Jakacki Diane Katherine	Pimenta Ricardo Medeiros
Dalmau Michelle	Jannidis Fotis	Piquette Kathryn E.
Dalziel Karin	Johnson Ian R.	Porter Dorothy Carr
Damala Areti	Jones Steven Edward	Posner Miriam
Dannaoui Elie	Jordanous Anna Katerina	Potvin Sarah
Davis Rebecca Frost	Kaislaniemi Samuli	Prescott Andrew John
	Kaplan Frédéric	Priani Ernesto
	Kelber Nathan Patrick	Puschmann Cornelius
	Kelleher Margaret	Pytlik Zillig Brian L.

Rahtz Sebastian
Rapp Andrea
Redwine Gabriela Gray
Rehbein Malte
Rehberger Dean
Rehm Georg
Reside Doug
Rhody Jason
Riddell Allen Beye
Ridge Mia
Ridolfo Jim
Robey David
Rochat Yannick
Rochester Eric
Rockwell Geoffrey
Rodríguez-Ortega Nuria
Roe Glenn H
Romanello Matteo
Roorda Dirk
Rosner Lisa
Ross Claire Stephanie
Roueché Charlotte
Roued-Cunliffe Henriette
Rudman Joseph
Ruecker Stan
Ruotolo Christine
Rybicki Jan
Sahle Patrick
Saklofske Jon
Salciute-Civiliene Gabriele
Sanderson Robert
Sanz Amelia
Sayers Jentery
Schlitz Stephanie
Schmidt Desmond
Schmidt Sara A.
Schöch Christof
Scholger Walter
Schreibman Susan
Schubert Charlotte
Seppänen Tapio
Shaw Ryan Benjamin
Shep Sydney
Shimoda Masahiro
Siemens Lynne
Siemens Raymond George
Simons Gary F.
Sinclair Stéfan
Singer Kate
Smithies James Dakin
Snyder Lisa M.
Sokół Małgorzata
Spence Paul Joseph
Sperberg-McQueen Michael
Spiro Lisa
Stadler Peter
Stokes Peter Anthony
Streiter Oliver
Sukovic Suzana
Sula Chris Alen
Suzuki Takafumi
Swanstrom Elizabeth Anne
Tabata Tomoji
Takseva Tatjana
Tasovac Toma
Teich Elke
Terras Melissa
Thaller Manfred
Theibault John Christopher
Thieberger Nick
Thomann Johannes
Todirascu Amalia
Tomabechi Toru
Tomasek Kathryn
Tomi_ Marijana
Tonelli Sara
Tonra Justin Emmet

Trettien Whitney
Trippel Thorsten
Tupman Charlotte
Underwood Ted
Valverde Mateos Ana
van Dalen-Oskam Karina
Van den Branden Ron
van den Herik H. J.
van Erp Marieke
van Hessen Arjan
van Hooland Seth
Van Zundert Joris Job
Varner Robert Stewart
Venecek John T.
VERDU RUIZ SILVIA
Verhoeven Deb
Vertan Cristina
Vetch Paul
Viana Vander
Viglianti Raffaele
Vogeler Georg
Walsh John
Walter Katherine L.
Wandl-Vogt Eveline
Warwick Claire
Watrall Ethan
Weidman Robert William
Weingart Scott
Wernimont Jacqueline D
Wieneke Lars
Wiesner Susan L.
Wilkins Matthew
Willett Perry
Winder William
Witt Andreas
Wittern Christian
Wolff Mark
Worthey Glen
Wrisley David Joseph
Wulfman Clifford Edward
Yin Xin
Zafrin Vika
Zeldenrust Douwe
Zöllner-Weber Amélie

Workshops

Are we there yet? Functionalities, synergies and pitfalls of major digital humanities infrastructures

Benardou, Agiatis

ATHENA R.C., Greece

Champion, Erik

Curtin University, Australia

Hughes, Lorna

University of Wales, United Kingdom

Chambers, Sally

Göttingen Centre for Digital Humanities, Germany

Dallas, Costis

University of Toronto, Canada

Dunning, Alastair

The Europeana Foundation, The Netherlands

Abstract

This pre-conference workshop aims to bring together leading scholars involved in major digital scholarly infrastructure projects such as Dariah, NeDiMAH, Europeana Cloud, ARIADNE, 3D ICONS, EHRI, DASISH, LARM, CLARIN, DiRT and DHCommons, in dialogue with practising digital humanists. Topics to be addressed include cultural heritage and digital media infrastructures, tools and services; the creation and curation of humanities digital resources; social and institutional issues of Digital Humanities infrastructures; and finally, lessons learnt from the role of digital humanities in pedagogy and academic curricula. It will provide an opportunity for humanists to find out about cutting edge developments on major digital infrastructure initiatives in Europe and beyond, and to make their views matter on future developments in this field.

The workshop aims to go beyond a description of project presentations. It will seek to provide an analytical framework that could contribute to a critical understanding of the current state of digital infrastructures vis-à-vis the potential of digital archives, tools and services for humanities scholarship, by addressing the following questions:

1. What are the objectives of each digital infrastructure project, and what are its intended users?
2. What are the functionalities and outcomes it aims to provide, and how do they serve the overarching goal of supporting and transforming humanities research?
3. To what extent were the needs of humanities researchers considered, and how is the digital humanities research community involved in the project?
4. Are there potential synergies, and actual collaboration, with other infrastructure projects? Conversely, are there any overlaps?
5. What are the main lessons learned from the life of the project so far? What are the pitfalls and potential failures, and what improvements could be achieved?

Workshop leaders

The workshop will be led by the following international team:

- **Dr Agiatis Benardou**, a member of the research staff of the Digital Curation Unit, IMIS-Athena Research Centre, Artemidos 6 & Epidavrou str., GR 151 25, Maroussi, Greece; email: a.benardou@dcu.gr, tel. +30 210 6875425. She initially worked for the Preparing Dariah - the Digital Research Infrastructure for the Arts and Humanities project. She also participates in the Greek Research Infrastructure Network for the Humanities (DYAS / Dariah-GR). She has also worked on EHRI (European Holocaust Research Infrastructure). She is leading a Work Package in the project "Europeana Cloud - Unlocking Europe's research via the Cloud". The objective of her work in this Project is Assessing

Researcher Needs in the Cloud and Ensuring Community Engagement.

- **Professor Erik Champion**, School of Media Culture & Creative Arts, Faculty of Humanities, Curtin University GPO Box U1987, Perth, WA 6845, Australia, email: erik.champion@curtin.edu.au. Champion writes on virtual heritage (Playing With The Past, Critical Gaming in the Digital Humanities), and on game-based learning for history and heritage (Game Mods: Design Theory and Criticism). He was Project leader of DIGIUMLAB Denmark, and VCC2 co-leader at Dariah. He is Professor of Cultural Visualisation at Curtin University, Australia.
- **Professor Lorna Hughes**, University of Wales Chair in Digital Collections, National Library of Wales, Aberystwyth SY23 3BU, UK; email: Lorna.hughes@llgc.org.uk. Hughes is the University of Wales Chair in Digital Collections, based at the National Library of Wales, where she leads a research programme in digital collections, researching and building projects that develop new digital content that addresses specific research or education needs, in partnership with academics and other key stakeholders in Wales and beyond. Her research focuses on the use of digital content in research, teaching, and community engagement. Her publications include the edited volumes Digital Collections: Use, Value and Impact (2011) and Virtual Representations of the Past (2008), and Digitizing Collections: Strategic Issues for the Information Manager (2003). She is Chair and (UK representative) on the ESF Network for Digital Methods in the Arts and Humanities (www.nedimah.eu), and the PI on a JISC-funded mass digitization initiative The Welsh Experience of the First World War (cymruww1.llgc.org.uk).
- **Sally Chambers**, Joint Secretary-General for Dariah-EU, the Digital Research Infrastructure for the Arts and Humanities, based in the Göttingen Centre for Digital Humanities, Heyne-Haus, Papendiek 16, D-37073 Göttingen, Germany; email: sally.chambers@phil.uni-goettingen.de. She has extensive experience in the field of digital libraries, as Digital Research Manager at The European Library (TEL), Online Library Manager at the University of London Library, and Liaison Officer of Electronic Access to Resources in Libraries (EARL). Her work focuses on interoperability, metadata and technical project coordination, and her previous projects include Europeana Libraries, a project to establish a sustainable library-domain aggregation service for Europe, and ARROW, a project to establish a rights information management infrastructure to facilitate digitisation in Europe. She has been actively involved in the European digital library community, including the European Library Automation Group (ELAG) and the Dublin Core Metadata Initiative. She is the editor of Catalogue 2.0: the future of the library catalogue, recently published by Facet Publishing.

Expected audience

The workshop will be of direct relevance to the conference topic of **Digital Cultural Empowerment**, as it aims to take stock and review critically the current state of play, and potential future developments, on digital make the voice of humanists heard on the subject of digital infrastructures for the arts and humanities, a potentially important empowering factor concerning digital humanities research. It is expected to be of interest both to those involved in digital research infrastructure work, and to digital humanists who may benefit from the use and contribute to shaping the plans for future developments of digital infrastructures, tools and services.

Structure and organisation of the workshop

The event will involve formal presentations by the organisers, and speakers who are developers and evaluators of current and future digital cultural heritage infrastructures.

The workshop organizers will provide summarized critical reviews to accepted submissions, and speakers will address generic leading questions from the organizers, and moderated dialogue with digital humanists present, so that we can provide

meaningful experience to the participants. We hope that this format will allow participants to discuss and understand where mistakes are made and how to evaluate and redesign and improve their own infrastructures.

CFP: The CFP will be published one week after workshop acceptance, and authors will have 4 weeks for submission and will receive replies within two weeks of submission deadline.

Speakers in the workshop will be selected from a call, addressed to partners of major digital infrastructures in the arts and humanities.

The program committee consists of the workshop leaders: Dr Agiatis Benardou, Professor Erik Champion, Professor Lorna Hughes, and Sally Chambers.

Background

In 2013 we ran a workshop on the subject of *Cultural Heritage Creative Tools and Archives* (<http://chcta.wordpress.com>), funded by the European Alliance of Digital Humanities, NeDiMAH, DIGHUMLAB, and the National Museum of Copenhagen (where it was hosted). The countries represented at this event included Austria, Belgium, Denmark, Germany, Greece, Lithuania, Norway, Romania, United Kingdom, and Canada. We were particularly pleased to have two invited speakers at this workshop who each have many decades of experience in cultural heritage infrastructures: Professor Julian Richards of York University, and Professor Sean Ross, Dean of the iSchool, University of Toronto.

From the two days of presentations we saw several issues reappearing in many of the presentations. Many EU institutes were duplicating (without realising) the work of others; there was little systematic evaluation of user needs (although several papers were exceptionally useful surveys of user needs before and after); and the lessons learnt from comparing the original aims and objectives with the final or posited audience needs were not always consistently followed through.

Program outline

For a half day, 3 hours plus breaks.

- 10 minute introduction and house rules.
- 120 minutes: 6 x 10 minute presentations (slides will have auto-timing), preceded by 3 minutes intro and summary of reviewers comments, presentation followed by 7 minute questions (so 20 minutes each presentation).
- 40 minutes open floor questions and facilitated discussion.
- 10 minute summation / wrap-up.

Total: 180 minutes; we will invite speakers to a restaurant dinner at own expense after the workshop.

Target audience: Digital humanists at large, including also archaeologists, heritage experts and historians, archivists, and those interested in cultural heritage infrastructures at European level. The 2013 CHCTA workshop had roughly 20 presentations and 40 in total attended.

Expected speakers: 8

Expected audience number: 30

Introducing the EpiDoc Collaborative: TEI XML and tools for encoding classical source texts

Bodard, Gabriel

gabriel.bodard@kcl.ac.uk
King's College London

Franzini, Greta

franzini@informatik.uni-leipzig.de
University of Leipzig, University College London

Stoyanova, Simona

simona.stoyanova@informatik.uni-leipzig.de

University of Leipzig, King's College London

Tupman, Charlotte

charlotte.tupman@kcl.ac.uk
King's College London

The EpiDoc Collaborative is a set of guidelines, schema and related tools for the encoding of epigraphic and other ancient text editions in TEI XML. The first EpiDoc Guidelines, published in 2000, arose jointly from work on Latin inscriptions by scholars at the University of North Carolina, and from work by the EAGLE Commission of the Association Internationale de l'Epigraphie Grecque et Latine. Since then, many major online editions of inscriptions have been published using EpiDoc, including the Inscriptions of Aphrodisias, Vindolanda Tablets Online, US Epigraphy Project, Inscriptions of Roman Tripolitania, Pandektis (Upper Macedonia, Aegean Thrace and Achaia), Roman Inscriptions of Britain, and now massive corpora such as the Duke Databank of Documentary Papyri, Datenbank zur jüdischen Grabsteinepigraphik and the EAGLE Europeana Project, make use of EpiDoc in their workflow.

Although conceived as a standard for digital epigraphy and papyrology, EpiDoc is also applicable outside these fields. It is well documented, and provides for numerous levels of transcription detail, while staying flexible enough to accommodate various structures of texts and editions. One of the main reasons behind the pending conversion of the Perseus Digital Library to EpiDoc-based TEI P5 is to ensure compatibility with the already existing epigraphic and papyrological corpora in EpiDoc. The EpiDoc standard (epidoc.sf.net), a specialization of the TEI originally developed for classical epigraphy and papyrology, is now being used for a broad range of texts which require deep and detailed markup as a result of the complex relationship between the text and the object on which it is written. Literary collections such as the Perseus Digital Library at Tufts and the Digital Fragmenta Historicorum Graecorum (DFHG) in Leipzig have recently adopted the EpiDoc schema for this very reason. In addition, the SoSOL EpiDoc editing interface has served as the basis of the Perseids platform, in development by the Perseus team, with further functionalities for editing and annotating texts online.

An average of two to three times per year, a week-long EpiDoc training workshop is held for trained epigraphists and papyrologists with no technical background. These workshops, run in London and elsewhere, regularly accommodate 20 or so students (at all levels from graduates to professors), and are always over-subscribed, sometimes with 50% or more applicants having to be turned away due to lack of space. These week-long events allow time for a basic introduction to XML, detailed discussion of epigraphic features (including text and edition structure) rendered in TEI, plenty of unstructured "workshop" time, and introduction to tools such as the Papyrological Editor and Example Stylesheets for rendering HTML editions. A one-day tutorial would obviously focus on a more limited subset of this material, necessarily covering it in less detail, but assuming a bit more technical experience as a starting point from a digital humanities audience.

This tutorial will benefit students, scholars and researchers as well as the general public interested in reinforcing their existing XML skills and learning to apply these to other materials and contexts (linked data).

The aim of this tutorial is to bring together international students, scholars and researchers already familiar with XML technologies to listen to their perspectives and needs, with a view of coupling research and practice by increasing their knowledge, enhance their skills and change their attitude towards the study and representation of a text. The tutorial will look at ways in which different classical source texts can be integrated and thus enriched via EpiDoc annotation, providing a testbed for larger and more complex projects. By the end of the tutorial, participants will be able to approach and analyse a text from an editorial and technical standpoint, with a view to expanding their knowledge to encompass a richer and wider variety of texts, join the EpiDoc community and pass on their skills.

Programme:

The day will begin with a short introduction to EpiDoc, its history and the theoretical basis of EpiDoc encoding. We will give an overview of the structure of a traditional epigraphic or papyrological edition, with reference to examples from databases and print editions, and show how TEI elements are mapped to the semantic distinctions and fields of such an edition. Some time will be given for practice. We will continue with further discussion of the Leiden Conventions (a set of rigorous and arbitrary sigla for encoding editorial features of transcribed text, in use in the classical discipline since 1931) and how we map TEI elements to the semantic features that they represent. The EpiDoc Guidelines and further examples will be shown, and more time given to practice. As a self-checking mechanism, students will be shown how to transform their EpiDoc XML files into an HTML page that represents the edition according to the conventions, using the example XSLT stylesheets provided by the EpiDoc collaborative.

The afternoon session will start with an introduction to the Papyrological Editor and the use of a tag-free interface. Participants will have the opportunity to enter a papyrological text into the database as an exercise. We will continue with a discussion on the principles of crosswalking; examples include EpiDoc to EDH and HGV to EpiDoc, as well as an example of EpiDoc's applicability to non-epigraphic material with the ongoing conversion of the Perseus Digital Library. Finally, we will explore the ways in which EpiDoc data can be linked with other resources and shared using RDF (Resource Description Framework). We will illustrate this using examples from resources such as Pelagios [pelagios-project.blogspot.co.uk], and will discuss ontologies that are relevant to materials encoded using EpiDoc, including Pleiades/Pelagios, SNAP:DRGN, and the Eagle Europeana project.

Participants are welcome to bring questions and problems arising from their own texts, and where feasible we will target these needs in our presentations and exercises.

Overview:

Morning:

- Introduction to EpiDoc, history and theory
- Structure of an epigraphic edition
- Leiden conventions
- How to transform your EpiDoc to a HTML page

Afternoon:

- Using the Papyrological Editor
- Principles of crosswalking
- Exposing EpiDoc as Linked Data
 - introduction to Linked Data/RDF
 - ontologies and vocabularies

Instructors:

Gabriel Bodard is a researcher in digital epigraphy at the Department of Digital Humanities at King's College London, and one of the lead authors of the EpiDoc Guidelines and toolset. He was on the Technical Council of the TEI for six years, and has been teaching EpiDoc workshops regularly since 2004.
 gabriel.bodard@kcl.ac.uk

Greta Franzini is a Classicist and a Digital Humanist. She works as a Research Associate for the Open Philology Project at the University of Leipzig. She is also undertaking a PhD under the supervision of Dr. Melissa Terras and Simon Mahony at the UCL Centre for Digital Humanities, where her research on electronic editing will ultimately inform the production of her own edition of the oldest surviving manuscript of St. Augustine's *De Civitate Dei*.
 franzini@informatik.uni-leipzig.de

Simona Stoyanova is a Classicist who specialises in epigraphy, and a Digital Humanist. She works as a Research Associate for the Open Philology Project at the University of Leipzig. She is also doing a PhD in Digital Classics at

King's College London. Her research is focused on the Greek and Latin epigraphic traditions in the mixed language population of the province of Thrace, with particular interest on palaeographical issues and their possible investigation through the DigiPal framework.

simona.stoyanova@informatik.uni-leipzig.de

Charlotte Tupman is a Research Associate at the Department of Digital Humanities, King's College London. Her background is in Classics and Epigraphy, having studied Classical Archaeology at King's College and completed her PhD on Roman funerary inscriptions at the University of Southampton. Since then she has worked on various ancient and modern text projects at the Department of Digital Humanities. She has been co-teaching the EpiDoc workshops since 2006 and is currently contributing to the latest version of the EpiDoc Guidelines. She also co-organises the regular UK Practical Epigraphy

charlotte.tupman@kcl.ac.uk

Building bridges between Lausanne and Leeds: Virtual Round Table Discussion on methods, recent solutions and new questions between scholars at the International Mediaeval Congress in Leeds and the Digital Humanities Congress in Lausanne

Bruhn, Kai-Christian

bruhn@fh-mainz.de

i3mainz - Institute for Spatial Information- and Surveying-Technology

Schwartz, Frithjof

frithjof.schwartz@adwmainz.de

Academy of Sciences and Literature Mainz

Motivation

In July 2014 the Digital Humanities Congress in Lausanne and the International Mediaeval Congress in Leeds are going to be celebrated contemporaneously.

Both congresses have a worldwide reputation and are widely recognised platforms for discussing current developments and showcasing new developments as well as recent research approaches in their respective disciplines.

Realizing the increasing impact of contributions from digital humanities in Leeds at this year's conference we felt the importance to foster more intense scientific exchange between the Digital Humanities and Mediaevalists. Thus, the idea was born to introduce a virtual round table discussion between the IMC Leeds and the DHC in Lausanne as a forum that could stimulate an interdisciplinary discussion at two places via live-streaming complemented by additional web-based participatory elements.

Topic

During the last decades Digital Humanities have developed out of their niche as an auxiliary scientific research field to an autonomous and well-founded discipline. The reasons for this development are as manifold as the diverging views on what this new discipline is covering, only about ten years after the alias Digital Humanities was coined.

The advance of information technology and the growing impact of the digital paradigm in everyday life has long reached science and humanities. The vast majority of research in the Humanities is transforming information into digital

representations and the web and other applications of the internet are used for the constitutive process of exchanging arguments and collating knowledge. At the same time, perception and acceptance of DH grows in the established Humanities disciplines not only because the passover to the world of the digital demands proper handling of the data but also because Digital Humanities succeed in proving its claim to contribute to generating new insights in a number of humanities-related fields.

However, there is no question that this process continues to gain momentum and that it is therefore necessary that both fields not just escort this change but rather actively shape its future path.

Therefore, we seek to deepen the already existing dialogue between Mediaeval Studies and Digital Humanities (in an innovative event) by initiating a well arranged Round Table Discussion on the impact and perspectives of studying the spatial aspects of Mediaeval written sources. We consider this being a key area of research in which both the Humanities and the Digital Humanities can easily be included and get in a dialogue from different points of views.

Agenda

In order to ensure a focused discussion that discloses the potentials and contributes to the diversification of DH approaches in Mediaeval studies, the format of the Round Table Discussion has to adopt not only the specific setting within two conferences but also its objective to stimulate interdisciplinary dialogue.

Three selected discussants at each panel in Leeds as well as in Lausanne will represent a cross section of approaches to the spatial aspects of mediaeval texts. The organisers will provide them with material on a well studied and prominent written source in advance of the event. The selection of the source takes into account its potential to serve as pars pro toto for a variety of ways approaching the spatial substance of historical sources. We will encourage a preparatory argument with all discussants to avoid misconceptions and to agree on a set of documents to be published in the run up of the conferences, open for public commentary and annotation. Based on the feedback, the discussants will agree on two main aspects to be discussed in the live panel. To introduce the audience, each member will give a brief introductory statement pinpointing to her or his understanding and perspective of the spatial information contained in the document (together ca. 20 min.). A first open panel will provide the opportunity for the audience to ask for specific explanation or to clarify misunderstandings (10 min.). In the subsequent two thematic blocks (20 min. each), the discussants will exchange views on the topics agreed upon. Before each group of discussants in Leeds as well as in Lausanne will close the session by summing up 'lessons learned' another 10 min. block will provide opportunity for enquiries or comments by the audience.

Two moderators will arrange for the coordination between the geographically separated sessions and the thematic focus of each block.

Technically we plan a livestream not only between the conferences but also in the web via services like e.g. ustream.tv offering a variety of possibilities for the web audience to track the discussion and to initiate follow-up debates in the social media.

Expected outcome

As an outcome of the round table discussion we envisage not only to clarify the different aspects and methods followed by scholars in the Digital Humanities in comparison to those of Mediaeval Studies but also to uncover new possibilities and fields of research initiated by an interdisciplinary discussion on a subject.

List of Participants

Lausanne:

- Prof. Dr. phil. Kai Christian Bruhn, i3mainz. Institute for Spatial Information and Surveying Technology FH Mainz - University of Applied Sciences, Germany
- Dorothy Porter MA., Curator, Digital Research Services at the Special Collections Center, University of Pennsylvania, USA
- Dr. Sarah Rees Jones, Department of History, University of York, England

Leeds:

- Prof. Dr. Sible de Blaauw, Faculty of Letters, Radboud University, Nijmegen, Netherlands (still uncertain)
- Dr. Kerstin Sailer, Bartlett School of Architecture, University College London, England
- Dr. Frithjof Schwartz, Akademie der Wissenschaften und der Literatur Mainz, Germany

A Collaborative, Indeterministic and partly Automatized Approach to Text Annotation

Bögel, Thomas

University of Heidelberg, Germany

Gius, Evelyn

evelyn.gius@uni-hamburg.de
University of Hamburg, Germany

Petris, Marco

University of Hamburg, Germany

Strötgen, Jannik

University of Heidelberg, Germany

1. Description

The webbased system CATMA (Computer Aided Text Markup and Analysis) was designed to address the interest essentially motivating human encounters with literature: hermeneutic, i.e., "meaning" oriented highorder interpretation. In the scholarly interpretation of literature we are not looking for the right answer, but for new, plausible and relevant answers. This requires a true hermeneutic markup as defined by Pietz (2010: paragraph 1):

By "hermeneutic" markup I mean markup that is deliberately interpretive. It is not limited to describing aspects or features of a text that can be formally defined and objectively verified. Instead, it is devoted to recording a scholar's or analyst's observations and conjectures in an openended way. As markup, it is capable of automated and semi-automated processing, so that it can be processed at scale and transformed into different representations. By means of a markup regimen perhaps peculiar to itself, a text will be exposed to further processing such as text analysis, visualization or rendition. Texts subjected to consistent interpretive methodologies, or different interpretive methodologies applied to the same text, can be compared. Rather than being devoted primarily to supporting data interchange and reuse – although these benefits would not be excluded – hermeneutic markup is focused on the presentation and explication of the interpretation it expresses.

CATMA has been developed to support McGann's (2004) openended, discontinuous, and nonhierarchical model of text-processing. Its nondeterministic approach to markup allows the user to express many different readings directly in markup. The system not only enables collaborative research but it is based on an approach to markup that transcends the limitations of lowlevel text description, too. CATMA supports highlevel semantic annotation through TEIcompliant, nondeterministic standoff markup and acknowledges the standard practice

in literary studies, i.e., a constant revision of interpretation (including one's own) that does not necessarily amount to falsification. Moreover, it enables users to switch ad hoc between text annotation and text analysis in either direction as well as recursively.

In 2013 in a joint project, heureCLÉA, two research teams (one computer scientists, the second narratologists) started to focus on an exemplary hermeneutic "use case": the decoding of temporal information in narratives, namely the automatic detection of temporal phenomena in literary narratives.

For this purpose, we developed an approach based on both manual annotation of narratological phenomena and the rule-based extraction and normalization of temporal expressions which are used as a starting point for machine learning. This project is still ongoing, but the automated annotation of temporal expressions and other linguistic features like POS (partofspeech) tagging and sentence detection, as well as tense annotations based on morphological analysis, have already been implemented in CATMA and can be used for a combined automatic and manual annotation of texts.

In our tutorial, we will introduce the core annotation and analysis functionalities of CATMA and show how they can be combined with the annotations provided automatically by HeidelTime and other components. Participants will have the opportunity of testing the tool in a hands-on session where they can annotate their own texts or annotate collaboratively a text we will provide. We would like to engage participants in a design critique of CATMA and its components and a general discussion about requirements for text analysis tools in their fields of interest, too.

2. Tutorial Instructors

All tutorial instructors come from the developing team of the heureCLÉA project. We have been presenting and teaching CATMA, HeidelTime and heureCLÉA on various national and international occasions in the last years. Two of us have included crucial aspects from heureCLÉA in their PhD research projects, too.

Thomas Bögel, Institute of Computer Science, Heidelberg University

Thomas studied computational linguistics and is currently working as a researcher and pursuing his PhD at the Institute of Computer Science at Heidelberg University. His research focuses on event extraction and timeline generation, as well as the development of machine learningbased systems for temporal relation extraction from narrative texts.

Evelyn Gius, Department of Languages, Literature and Media, Faculty of the Humanities, University of Hamburg

Evelyn has been trained as a computational linguist and is now working in the field of literary computing as a researcher and lecturer. For her PhD project she has explored with CATMA the benefits of applying narratological categories from literary studies to the analysis of narrations of reallife labor conflicts.

Marco Petris, Department of Languages, Literature and Media, Faculty of the Humanities, University of Hamburg

Marco is a computer scientist with a strong affinity for the humanities and has been engaged in the creation of CATMA from the very beginning. As a research developer he is involved in all aspects of the design and implementation of tools for the Digital Humanities.

Jannik Strötgen. Institute of Computer Science, Heidelberg University

Jannik studied computational linguistics and economics at Heidelberg University before he joined the Institute of Computer Science as researcher and PhD student. His research focuses on temporal and geographic information extraction and retrieval, and he is the main developer of the widelyused, multilingual, crossdomain temporal tagger HeidelTime, which achieved the best results for the task of temporal tagging at TempEval2010 (English) and TempEval2013

Contact address

Evelyn Gius
Universität Hamburg
Department of Languages, Literature and Media Institut für Germanistik

VonMellePark 6
20146 Hamburg
Tel + 49 40 42838 6942
Fax + 49 40 42838 3553 evelyn.gius@uni-hamburg.de

3. Target audiences and number of participants

The primary users of CATMA are literary scholars, and graduate and undergraduate students of Literary Studies. Nevertheless, this tutorial is likely to be of interest also to:

- humanities scholars of ALL fields concerned with text analysis (with and without experience in digital text analysis)
- software developers in the humanities interested in non-deterministic text analysis and automated annotation

Expected number of participants: We can accommodate up to 25 participants.

4. Special requirements

Participants will be asked to bring their own laptops. We will need internet access for all participants and a screen projector.

5. Outline of the tutorial

The tutorial is designed as a 3,5 hours tutorial, including a break of approx. 30 minutes. Provisional format:

- introduction to CATMA (10 min)
- the CATMA approach to markup: indeterministic and collaborative markup functionalities (20 min)
- automated tagging of temporal expressions provided by HeidelTime and other linguistic annotations by the UIMA-pipeline (30 min.)
- hands-on session: annotating texts (30 min) (break: 30 min)
- hands-on session: annotating texts (60 min)
- the heureCLÉA approach to narratological phenomena of time (15 min)
- wrap up discussion (15 min)

References

www.catma.de (last seen 20140217)

Piez, Wendell (2010), *Towards Hermeneutic Markup: An architectural outline*, King's College, DH 2010, London. Available from: <http://dh2010.uchc.kcl.ac.uk/academic-programme/abstracts/papers/html/ab743.html> (last seen 2014-0217).

We define this distinction as follows: description cannot tolerate ambiguity, whereas an interpretation is an interpretation if and only if at least one alternative to it exists. Note that alternative interpretations are not subject to formal restrictions of binary logic: they can affirm, complement or contradict one another. In short, interpretations are of a probabilistic nature and highly context dependent.

heureCLÉA is a BMBFfunded eHumanities project run jointly by the University of Hamburg and Heidelberg University since the beginning of 2013 (cf. www.heureclea.de, last seen 201402-17).

cf. the accepted paper by **Janina Jacke and Jan Christoph Meister**: *Pushing Back the Boundary of Interpretation: Concept, Practice and Relevance of a Digital Heuristic*

Sharing digital arts and humanities knowledge: Dariah as an open space for dialogue

Chambers, Sally

sally.chambers@phil.uni-goettingen.de
Göttingen Centre for Digital Humanities

Schmunk, Stefan

schmunk@SUB.UNI-GOETTINGEN.DE
Göttingen State and University Library

'I realized that sharing information is one of the most important roles of a digital scholar'
**(Irina Savinetskaya, CENDARI Research Fellow 2013,
Göttingen Centre for Digital Humanities)**

1. Outline

The aim of this pre-conference workshop is to bring together arts and humanities researchers and research infrastructure professionals in an open space for dialogue. The development of coherent research infrastructures in the arts and humanities in Europe is currently in its infancy. With the launch of the Horizon 2020 work programme on European research infrastructures¹ in December 2013, now is an ideal moment for researchers and infrastructure specialists to come together to shape the future of digitally-based research in the arts and humanities.

It is essential that the network of services offered by research infrastructures are developed by researchers, for researchers. Research infrastructure providers therefore need to ensure that they work closely with arts and humanities research communities to understand how they work, what challenges they face and offer solutions using appropriate digital technologies with a view to making their day-to-day research easier. Similarly, it is important that humanities researchers have a clear insight into the technological possibilities that research infrastructures could offer and understand the issues that research infrastructure specialists need to take into account e.g. security aspects, service level agreements, costs etc.

The controversy around research infrastructures in the humanities was one of the issues addressed during the Cologne Dialogue on Digital Humanities 2012². As a result, the organisers of this workshop would like to try using 'Open Space Technology'³, which has already been successfully proven in the digital humanities context with THATCamp (The Technology and Humanities Camp)⁴. Open Space Technology is 'effective in situations where a diverse group of people must deal with complex and potentially conflicting material in innovative and productive ways'⁵. This would therefore seem like a good methodology to use for this workshop.

The outcomes of the workshop are to increase understanding and encourage dialogue between arts and humanities researchers and research infrastructure professionals, with a view to jointly identifying concrete requirements for future developments.

The vision for Dariah (Digital Infrastructure for the Arts and Humanities)⁶ is to offer a portfolio of infrastructure-orientated activities centred around research communities across the broad spectrum of the arts and humanities. This pre-conference workshop at DH2014 is intended to be just one step in the process of making this happen and bridging the gap between humanities researchers and research infrastructure specialists.

2. Target audience

Researchers from across the arts and humanities, particularly those who use or are interested using digital methods in their research. Research Infrastructure specialists including librarians, archivists, curators, computer scientists, programmers, information scientists and research data managers.

3. Workshop schedule

We have invited an international 'virtual' Programme Committee who has agreed to assist the organisers in preparing this one-day workshop. At this stage, we anticipate that the workshop will use a blended approach of techniques, e.g. formal presentations, poster sessions and Open Space Technology, to help ensure maximum engagement of all

participants. An outline schedule is included below as a basis for the Programme Committee to build on:

09.00-09.15	Welcome & Introduction
09.15-10.30	Dialogue I: Researchers' interests
10:30-11:00	Coffee break
11:00-12:30	Dialogue II: Key note (15-20 min.) on the relation of research and infrastructure - an overview of the landscape with following discussion
12:30-13:30	Lunch break (and poster-session)
13:30-15:00	Dialogue III: Infrastructural offer
15.00-15.30	Coffee Break
15.30-16.30	Concluding Discussion
	"Shaping the future of digital arts and humanities research in Europe"

4. Workshop Outcomes

During the workshop we will use collaborative tools, such as Twitter (Hashtag: #DARIAHdialogue), Etherpad and a wiki-space / blog to develop a collaboratively authored digital record of the workshop. After the event, this would be curated and published as an open access, community-reviewed publication sustainably archived in a trusted digital repository.

5. Number of participants

Based on several recent workshops conducted by DARIAH, we suggest a maximum number of 40 participants to allow for dialogue and close interaction.

6. Workshop organisers**Sally Chambers (DARIAH-EU)**

Sally.Chambers@phil.uni-goettingen.de
00 49 551 39 20476

is a digital librarian, who leads the DARIAH-EU Coordination Office based in the Göttingen Centre for Digital Humanities, Germany. Before joining DARIAH-EU, Sally worked for The European Library, focusing on interoperability, metadata and technical project coordination. Her academic background is in literature, cultural studies and information services management.

Stefan Schmunk (DARIAH-DE)

schmunk@sub.uni-goettingen.de
00 49 551 39 - 20326

is a digital historian at the Goettingen State and University Library and currently working as project coordinator for DARIAH-DE and leading the cluster "Research Data Collections". His main focus is on encouraging the research-oriented interaction and needs of digital academics and IT-specialists and enabling the methodical discourse in the field of digital history.

7. Programme Committee

The people who have agreed to participate in the workshop Programme Committee are:

Aurélien Berra is an Assistant Professor at Paris-Ouest University, where he teaches Rhetoric and Ancient Greek literature. He is also in charge of the seminar "Digital Humanities" at EHESS and involved in DARIAH, HASTEC and Hypothèses. His special interest in digital textual scholarship stems from his work as a classical philologist.

Arianna Betti is Professor of Philosophy of Language at the University of Amsterdam. After studying historical and systematic aspects of ideas such as axiom, truth and fact (Against facts, MIT Press, 2014), she is now trying to trace the development of ideas such as these with computational techniques.

Mirjam Blümm is a digital librarian at the Goettingen State and University Library and currently working as project coordinator for TextGrid and DARIAH-DE. Her main focus

is on encouraging the discourse between digital humanities researchers, IT-specialists and information scientists. She is also active in education and teaching DH modules at Würzburg University.

Franz Fischer holds a position as research associate at the Cologne Center for eHumanities (CCeH), University of Cologne and is currently coordinating the EU funded Marie Curie action "DiXIT – Digital Scholarly Editions Initial Training Network". He is a founding member of the Institute for Documentology and Scholarly Editing (IDE).

Emiliano Degl' Innocenti holds a Ph.D. in Antique, Medieval and Renaissance Studies and works in Italian research institute (SISMEL - Fondazione Ezio Franceschini) as Head of the Computing in the Humanities Department. He is involved in DARIAH-IT, coordinator of the Medieval Prototype for the CENDARI project and invited expert for COST Action IS1005, Medieval Europe - Medieval Cultures and Technological Resources.

Franco Niccolucci is the coordinator of ARIADNE, a research infrastructure for digital archaeology, at the VAST-LAB of PIN, Italy. A former professor at the University of Florence, he directed the Science and Technology in Archaeology Research Center in Cyprus. His research interests focus on digital archaeology and its semantic foundations.

Ruth Reiche holds an M.A. in Art Education, Art History and Philosophy (LMU München). In 2009 she started working on her PhD thesis about storytelling in multiscreen installation (advisors: Prof. Stemmerich - FU Berlin; Fabienne Liptay - UZH Zürich). Since 2011 she has been working as a research associate at TU Darmstadt in DARIAH-DE, focusing on digital research methods. She is particularly interested in Contemporary Art, Digital Art History and Data Visualization.

Eveline Wandl-Vogt is Senior Scientist (Austrian Academy of Sciences [AAS]) and Research Manager of European Projects, e.g. WG-Chair @ COST IS 1305, Co-Chair of the Virtual Competency Centre on elInfrastructure @ DARIAH-EU. She is a data analyst, working on lexicography, geolinguistics, digital standards and encoding and has a strong practical expertise in DH and Interdisciplinary Humanities. She is currently working on building up the Austrian Centre for Digital Humanities (ACDH) with her Colleagues at the AAS.

Lars Wieneke holds a PhD in Engineering from the Bauhaus-University Weimar, Germany and joined the Centre Virtuel de la Connaissance sur l'Europe (CVCE) in 2011 where he now works as a researcher in the Digital Humanities Lab. Lars is a work-package leader in the FP7-IST funded research project CUBRIK, a member of the NeDIMAH and DHBenelux steering committee and has been a co-chair of a Europeana task force on User-Generated Content.

Hacking with the TEI

Ciula, Arianna

University of Roehampton, United Kingdom

Czmiel, Alexander

Berlin-Brandenburg Academy of Sciences and Humanities, Germany

Mylonas, Elli

Brown University, United States of America

Rahtz, Sebastian

Oxford University, United Kingdom

Cummings, James

Oxford University, United Kingdom

Syd, Bauman

Northeastern University, United States of America

Description

Digital humanists, electronic publishers, and many others use the Text Encoding Initiative (TEI) Guidelines to mark up electronic texts, and over time have created a critical mass of

XML — some conforming to subsets of the TEI Guidelines, some to individual customizations; in some cases intricate and dense, in others lean and expedient; some enriched with extensive external metadata, others with details marked explicitly in the text. The fruits of this labor are most often destined for display online or on paper (!), indexing, and more rarely, visualisation. Techniques of processing this markup beyond display and indexing are less wellunderstood and not accessible to the broad community of users, however, and programmers sometimes regard TEI XML as overcomplex and hard to process.

Our intent in this hackathon/workshop is to further a practice-based enhancement of input, output, and processing of TEI XML by:

- developing understanding and expertise of how to process TEI XML and get the best results from highlymarked up text resources;
- experimenting with a wider range of applications, particularly in visualisation.

Participants

The workshop will attract reasonably experienced DH practitioners who have not hitherto experimented with TEI XML, and those who have already been using TEI and developing TEI tools. The workshop will aim at engaging participants in the development of tools or techniques which are based on community needs and will be publicized to eager users. They will be stimulated by others working on similar problems, and the results of the hands on collaboration at the workshop will be promoted to the DH community at large.

Participants will work together in a welcoming and participatory environment, with TEI and other technical experts (on hand and available remotely) to explain the intricacies of the TEI and its applications. Sample texts will be provided where necessary. Participants will work together in small teams.

Challenges

In addition to the posting of the workshop on the DH 2014 website, we will issue a call for participation by March 20th at the latest (shortly after the notice of this proposal acceptance). We will ask applicants to describe their skills and interests, and to propose a project for the hackathon. Once participants have been selected, we will set up a collaborative platform in the form of a wiki, and participants will be encouraged to comment on the proposed tasks and to propose new ones to tackle together during the workshop. A few projects will be selected by the group before the event. Tasks selected will be tuned to a range of skill sets. We will also request suggestions from the TEI community via its mailing lists.

Possible challenges might include but are not limited to:

- mining a large corpus of texts for some data facet and visualising the results
- rendering complex markup in an innovative and playful way
- writing input or output filters for existing bits of software
- extending existing TEI software to take advantage of external resources such as Zotero
- adding a TEI mode to a web editor
- applying visualisation to TEI documents or schemas (e.g. visualizing the TEI conceptual model)

Outline

Length: 1 Day

Morning:

- 09:00–10:00 introduction and coffee, finalise groups and challenges.
- 10:00–12:30 groups start work (break out sessions)
- Lunch:
- 12:30–13:30 lunch and groups report on work so far
- Afternoon:

- 13:30–16:00 group work continues (break out sessions)
- 16:00–17:30 regroup, report back and show work, plan for next steps, evaluation form

Follow up

Participants will have the option of applying for a \$1000 grant from the Text Encoding Initiative to allow them to finish their work and make it available to others. Details for this competition will be provided after the workshop has taken place.

Organisers and TEI Experts

The organisers of this proposal all have extensive experience in leading and coordinating hands on workshops focused on TEI or related theories and technologies. They will be available during the workshop together with other TEI and DH experts who will be attending DH 2014 and confirmed interest to take part in this initiative should it be accepted by the programme committee:

- Syd Bauman (Northeastern University, US TEI council member, present) expertise: TEI (including ODD), XSLT, RELAX NG, Schematron, Perl, bash
- Hugh Cayless (Duke Collaboratory for Classics Computing, Duke University, US workshop organiser and TEI council member, remote) expertise: document analysis, modelling, EpiDoc, XSLT, XML processing in various programming languages
- Arianna Ciula (University of Roehampton, UK workshop organiser and TEI board member, present) expertise: document analysis, modelling, hybrid publications, TEI integration with semantic models
- James Cummings (IT Services, University of Oxford, UK TEI Technical Council Chair,) expertise: TEI, XSLT, XQuery, basic jQuery
- Alexander Czmiel (BerlinBrandenburg Academy of Sciences and Humanities, Germany) expertise: TEI, XSLT, XQuery, eXist, Digital Editions
- Elli Mylonas (Center for Digital Scholarship, Brown University workshop organiser and TEI council member, present) - expertise: document analysis, modelling, experience with TEI users/encoder training, TEI workflows
- Sebastian Rahtz (IT Services, University of Oxford, UK workshop organiser and TEI council member, present) expertise: processing TEI ODD specifications, ePub generation, and programming in XSL

“What's your method?” Building an ontology for digital research methods in the arts and humanities

Constantopoulos, Panos

Dallas, Costis

Hughes , Lorna

Thaller, Manfred

Background

Digital research methods in the arts and humanities have been the focus of important systematic work for more than a decade, taking the form of a digital (computational) methods taxonomy developed by AHDS in the UK, and then expanded and reused in several digital humanities initiatives internationally. The taxonomy was adopted as the conceptual structure for a series of ICT Guides for digital arts and humanities in the UK, the arts-humanities.net portal of digital humanities projects, tools, methods, expert centres,

researchers, and papers, and the Database of Research and Projects in Ireland (DRAPler, Digital Humanities Observatory). It was recently refined by DARIAH-DE in collaboration with the Bamboo DiRT project. It is now the focus of a joint project by the Network for Digital Methods in the Arts and Humanities (NEDIMAH) and DARIAH-EU, which aims to develop an ontology of digital research methods in the arts and humanities: a formal conceptualization of digital research methods and their context of scholarly use, which can be used to adequately represent the domain of arts and humanities scholarly practice in the digital age. These include equally methods of information seeking, use and modification of digital resources used in scholarly work, and computational methods used by humanities scholars in all phases of the scholarly research lifecycle, from the generation of a research question or topic, to the representation, visualisation and analysis of research data and sources, and to scholarly publication and communication.

Workshop structure

This workshop aims to engage participants in the theory and practice of developing an ontology for digital research methods in the arts and humanities, through an interactive modeling activity, led by a team of experts in humanities digital research methods and ontology building. Participants will be introduced to the background and state-of-the-art regarding the domain of digital computational methods in the arts and humanities, as well as to the scholarly processes and “research primitives” associated with digital methods, tools and services. They will also be provided with a graduated introduction to ontologies, to the main concepts and techniques involved in developing an ontology, and to a conceptual model of research scholarly activity suitable for representing the application of digital methods for arts and humanities research. They will then be invited to share and discuss short informal accounts of their own scholarly work, focusing on the use of digital sources, tools, and services, which will provide the basis for a conceptual analysis and ontology building hands-on exercise, based on identifying conceptual relationships and integrating the insights derived by individual digital research experiences into a shared conceptualisation under the guidance of workshop leaders.

Workshop participants

The workshop will be of interest to both advanced digital humanities scholars and to digitally-enabled humanities researchers, or those not currently using digital tools and methods but interested to do so in the foreseeable future. It will be open to those occupied with the study of textual and visual resources, material and intangible cultural heritage, quantitative and qualitative modes of analysis, and a variety of epistemological stances within digital humanities. In addition, the workshop may be attractive to STS scholars interested in understanding scholarly practice, as well as to computer scientists, information scientists and others interested in the relationship between digital humanities and digital infrastructures. Participants will not be required to have prior knowledge in the field of ontology engineering, but should be familiar with particular research methods in the arts and humanities, interested in reflexive analysis of humanities research practices, and prepared to engage, under the guidance of workshop leaders, with formal methodologies of ontology modelling.

Benefits for participants include the opportunity to share experiences, reflect critically on, and discuss the methods employed in digitally-based humanities research; to conceptualise specific digital research methods in the context of particular kinds of research, types of resources, and digital tools and services; and, to enhance their understanding of and get acquainted with basic ontology building techniques useful in the domain of digital humanities.

Workshop leaders

The workshop will be led by an international team with expert knowledge in the field of digital humanities methods and cultural ontologies, currently involved in a major joint research project of building a digital methods ontology for the arts and humanities under the auspices of NeDiMAH – Network for Digital Methods in the Arts and Humanities, and DARIAH-EU – Digital Research Infrastructure for the Arts and Humanities in Europe:

Prof. Panos Constantopoulos is Professor and Dean, Faculty of Information Sciences, Athens University of Economics and Business, and Director of the Digital Curation Unit, IMIS-Athena R.C. (Artemidos 6 & Epidavrou, Maroussi GR 151 25, Greece; p.constantopoulos@hua.gr). He has previously been Professor at the Department of Computer Science, University of Crete, where he has also served as Department Chairman and Director of Graduate Studies. He has founded and led for twelve years the Information Systems Laboratory and the Centre for Cultural Informatics at the Institute of Computer Science, Foundation for Research and Technology - Hellas. He holds a Diploma in Electrical and Mechanical Engineering from the National Technical University of Athens, a Master of Science in Electrical Engineering from Carnegie-Mellon University, and a Doctor of Science in Operations Research from the Massachusetts Institute of Technology. His research interests are in information systems, knowledge representation and conceptual modelling, ontology engineering, semantic information access, information design, decision support and knowledge management systems, cultural informatics and digital libraries. He has been principal investigator in 35 national and international competitive research projects, in 3 of which he was project co-ordinator. He has about 90 articles published in scientific journals, the proceedings of international scientific conferences, or as chapters in books.

Prof. Costis Dallas is Director of Museum Studies and Associate Professor at the Faculty of Information, University of Toronto (140, St George str, Toronto ON M5S 3G6; costis.dallas@utoronto.ca). His recent and current work as Research Fellow of the Digital Curation Unit-IMIS, Athena Research Centre, and co-principal investigator in the CARARE, LoCloud, Europeana Cloud and ARIADNE projects, and as Chair of VCC2 Task 2 – “Understanding and expanding scholarly practice” of DARIAH-EU, focuses on understanding knowledge practices and digital research methods in the field of cultural heritage and humanities scholarship, on knowledge representation of material culture, and on the specifications of curation-aware cultural heritage metadata repositories. He is currently engaged in developing a theoretical framework for the digital curation of “thing cultures”, integrating historical approaches to the representation and study of cultural objects with methodologies, infrastructures and environments intended for the management, preservation and use of digital information. He holds a DPhil degree in Classical archaeology from the University of Oxford.

Prof. Lorna Hughes is University of Wales Chair in Digital Collections, National Library of Wales (Aberystwyth SY23 3BU, UK; Lorna.hughes@llgc.org.uk). At the National Library of Wales, where she leads a research programme in digital collections, researching and building projects that develop new digital content that addresses specific research or education needs, in partnership with academics and other key stakeholders in Wales and beyond. Her research focuses on the use of digital content in research, teaching, and community engagement. Her publications include the edited volumes *Digital Collections: Use, Value and Impact* (2011) and *Virtual Representations of the Past* (2008), and *Digitizing Collections: Strategic Issues for the Information Manager* (2003). She is Chair and (UK representative) on the ESF Network for Digital Methods in the Arts and Humanities (www.nedimah.eu), and the PI on a JISC-funded mass digitization initiative The Welsh Experience of the First World War (cymruww1.llgc.org.uk).

Prof. Manfred Thaller is Professor of Computer Science for the Humanities, University of Cologne (Kerpener Str. 30, Köln; manfred.thaller@uni-koeln.de). His current research interests are in the theory of a Computer Science for the Humanities; non-relational data models; and the relationship between Markup Languages and DBMS. He has been involved in a large number of digital humanities research projects, including,

recently: the Digital Manuscript Library of Cologne (CEEC), the German Art Historical Decentralised Imager Archive (Prometheus), the Frankfurt digital library, and the Manuscript server project Duderstadt. He has a long career of academic research, publishing and teaching in modern history, empirical sociology and digital humanities, including appointments at the Institute of Advanced Studies in Vienna, the Max-Planck-Institut for History at Göttingen, and the University of Bergen. He was responsible for the design and implementation of a general data base oriented programming system for history (CLIO/κλειώ), while also working on developing a general methodology of historical computer science.

Annotation Studio: an open-source, collaborative multimedia online note-taking tool for humanities teaching and learning

Fendt, Kurt

fendt@mit.edu
MIT, United States of America

Folsom, Jamie

jfolsom@mit.edu
MIT, United States of America

Schnepper, Rachel

schnep@mit.edu
MIT, United States of America

Andrew, Liam

landrew@mit.edu
MIT, United States of America

Description

Annotation Studio is a collaborative online annotation tool developed at Hyperstudio, Center for Digital Humanities at MIT, for teaching and learning in the humanities. It is funded by the National Endowment for the Humanities and is one of several annotation tools that incorporate The Annotator from the Open Knowledge Foundation (annotatorjs.org), thereby benefiting from and participating in the highly collaborative community around that code library.

This is a half-day (3-hour) workshop for educators, administrators, librarians, developers and technologists interested in exploring what Annotation Studio has to offer. The content and format has been refined based on feedback from workshops delivered over several years. All participants will come away with an understanding of the tool and what it can do.

In addition, we aim to convey the adaptability of Annotation Studio and to illustrate how individuals and institutions can adopt it, integrate it into humanities teaching and learning, deploy it, extend it for specific use cases, and how to contribute back to the growing community of practitioners.

We'll open with a walkthrough of the tool and its functionality, then quickly move to structured hands-on work, followed by focused discussions and a final Q&A and networking opportunity. In the hands-on session, all participants will take on the roles of readers, writers, and instructors. Through this exchange of roles, all will gain a comprehensive understanding of the functionality of the tool. Participants will then choose one of two breakout sessions—one on pedagogy and theory or one on development and deployment—to focus on their more particular areas of interest.

- Pedagogy, curriculum, and theory. Participants in this session, led by Hyperstudio Director Kurt Fendt and Communications Officer Rachel Schnepper, will gain insights from instructors into how the tool has been used in humanities teaching, examples of assignments and ways to integrate the tool into the curriculum, and discuss possible

- new applications of annotation in their own field. Forms of assessment will be discussed as well.
- In the development, deployment, and administration session, led by Hyperstudio Web Applications Developer Jamie Folsom and Research Assistant Liam Andrew, participants will look at how they can add features to the tool, deploy it on their own infrastructure, and contribute their feature changes back to the project for use by others. We will invite technologists attending the conference who have adopted Annotation Studio to share their experience in this session.

In both sessions, we will introduce an online support forum we are using to facilitate conversations across time and distance, and invite all participants to join that conversation.

Outline

Overview

- Goals of the workshop
- Tour of the tool
- Resources including support forum and user manual

Hands-on

Reader

- Read and annotate
- Search and organize annotations
- Compose and get feedback

Instructor

- Add and manage documents
- Create groups and organize tasks
- View and assess readers' work

Break out sessions

Pedagogy, Curriculum, and Theory

- Sample syllabi and assignments
- Discussion of classroom integration
- Assessment
- Use cases by participants

Development, Deployment, and Administration

- Run an instance of the application
- Customize the application
- Contribute to the project

Debrief

- Discussion, feedback, support
- Q&A / Brainstorm (discussion)
- Networking (marketplace)

Target Audience

The target audiences for this workshop include: students, instructors, developers, and administrators in the humanities, in both formal and informal settings. In the past, this workshop has attracted between 30 and 40 participants from diverse fields and institutions. All participants should bring a laptop equipped with Wi-Fi with the latest Google Chrome, Safari, or Mozilla Firefox web browser installed prior to the workshop. The application does not work offline and does not work in Internet Explorer.

Core content

Annotation Studio is an easy-to-use platform for collaborative note taking and commenting with a special focus on education. This workshop provides an introduction to the application (30 minutes), and hands-on practice in different roles (60 minutes). Two breakout sessions will address topics of interest to specific audiences in more depth (60 minutes). The workshop wraps up with 30 minutes of question/answer, feedback, and networking.

Length and format

Part 1: Intro, (30 minutes)

- Part 2: Hands-on (60 minutes)
- Part 3: Breakouts (60 minutes)
- Part 4: Debrief (30 minutes)

Workshop Organizers

Kurt Fendt, Executive Director, MIT Hyperstudio

Dr. Kurt Fendt is Principal Research Associate in Comparative Media Studies and Executive Director of HyperStudio – Digital Humanities at MIT. He teaches a range of upper-level German Studies courses in Foreign Languages and Literatures. Fendt has held Visiting Professorships at the University of Cologne, the Technical University of Aachen (both Germany), and the University of Klagenfurt, Austria; in 2001 he was Visiting Scientist at the Fraunhofer Institute in Sankt Augustin, Germany. He is co-Principal Investigator of Annotation Studio, an NEH-funded web application for multimedia annotation in humanities education. Since 2005, he has been organizing the MIT European Short Film Festival. Before coming to MIT in 1993, Fendt was Assistant Professor in the Department of Applied Linguistics at the University of Bern in Switzerland, where he established the Media Learning Center for the Humanities and earned his Ph.D. in modern German literature with a thesis on hypertext and text theory in 1993 after having completed his MA at the Ludwig-Maximilians-University in Munich, Germany.

Jamie Folsom, Lead Web Applications Developer, MIT Hyperstudio

Jamie builds tools to support teaching and research in the humanities. He participates in all aspects of the lab's work, from consulting with faculty and collaborating with partners, to developing and deploying web apps and services, to presenting the lab's work and training people on the use of its tools at conferences. He has extensive experience teaching with and about technology, managing technology projects, and building web sites and applications.

He is particularly interested in the technical challenges peculiar to Digital Humanities Centers: how to conceive, design and develop tools highly tailored to specific research and instructional goals, while remaining agile enough to serve a range of disciplines and fields.

He holds an AB in French from Vassar College and a Master's Degree in Technology in Education from Harvard University, and has been a teacher, a technology trainer and manager, and a web applications developer for 20 years. He is from Boston, Massachusetts.

Rachel Schnepper, Communications Officer, MIT Hyperstudio

As Communications Officer, Rachel brings over ten years of higher education experience with her to HyperStudio. Prior to working at HyperStudio, Rachel taught at Rutgers University, Princeton University, DePaul University, and Washington and Lee University. Accordingly, Rachel is intimately familiar with the needs of faculty and is committed to helping them integrate digital humanities tools into their research and teaching.

Rachel earned her PhD in early modern European history in 2010 from Rutgers University. The Andrew W. Mellon Foundation and the North American Conference on British Studies have supported her research, which focuses on media transformations in the seventeenth century English Atlantic.

Liam Andrew, Research Assistant, MIT Hyperstudio

Liam Andrew graduated from Yale University in 2008, where he studied the advent of sound recording and its influence on modern language, literature and music. After stints as a book indexer, French-to-English translator, archivist, and English teacher abroad, he dove into programming and emerged as a software engineer for Delve, a newsreader and aggregator that helps organizations find and share important reads. As a graduate student in MIT's Comparative Media Studies program, his research interests lie at the intersection of sound and text on one hand, and classification and recommendation systems on the other. He is also a sound designer for theater and multi-instrumentalist in Dinowalrus .

Contact

Kurt Fendt, Executive Director, MIT HyperStudio
MIT HyperStudio
MIT Room 16-635
77 Massachusetts Avenue Cambridge, MA 02139, USA
phone: 617-253-4312
skype: kendt
email: fendt@mit.edu
twitter: @fendt

Jamie Folsom, Lead Web Applications Developer
MIT HyperStudio
MIT Room 16-635
77 Massachusetts Avenue Cambridge, MA 02139, USA
mobile: 617-669-0852
skype: jamiefolsom
email: jfolsom@mit.edu
twitter: @jamiefolsom

Rachel Schnepper, Communications Officer
MIT HyperStudio
MIT Room 16-635
77 Massachusetts Avenue Cambridge, MA 02139, USA
phone: 617-324-0102
email: rschnepp@mit.edu

Liam Andrew, Research Assistant
MIT HyperStudio
MIT Room 16-635
77 Massachusetts Avenue Cambridge, MA 02139, USA
email: landrew@mit.edu

GIS in the Digital Humanities: An introductory workshop

Gregory, Ian

I.Gregory@lancaster.ac.uk
Lancaster University, United Kingdom

Barker, Elton

Elton.Barker@open.ac.uk
The Open University, United Kingdom

Lang, Anouk

anouk.lang@strath.ac.uk
University of Strathclyde

Description:

Geographical Information Systems (GIS) are becoming increasingly used by historians, archaeologists, literary scholars, classicists and others with an interest in humanities geographies. To date, however, adoption of the technology has been hampered by a lack of understanding of what GIS is and what it has to offer to these disciplines. This introductory workshop, will provide a basic introduction to GIS both as an approach to academic study and as a technology. Its key aims are: To establish why the use of GIS is important to the humanities; to stress the key abilities offered by GIS, particularly the capacity to integrate, analyse and visualise data from many different types of sources; to show the pitfalls associated with GIS and thus encourage a more informed and subtle understanding of the technology; and, to provide a basic overview of GIS software and data. The workshop builds on a format that has been used successfully in the past, combining an overview of what GIS and what it has to offer to the humanities, contrasting case studies of its actual use, and a roundtable discussion. Please note that it does not provide hands-on software training, this can be gained from longer courses such as the *Lancaster Summer Schools in Interdisciplinary Digital Methods* at Lancaster University (<http://ucrel.lancs.ac.uk/summerschool>), and the Digital Humanities Summer Institute (DHSI) at the University of Victoria (<http://www.dhsi.org>)

The workshop will be split into three sessions of approximately one hour each as shown below assuming a

1:00pm start time. As discussed below, these timings are flexible and could be extended.

- 1:00-1:15: Welcome and Introductions
- 1:15-2:15: Session 1: Fundamentals of GIS from a humanities perspective
- 2:15-3:15: Session 2: Case studies on the use of GIS in the humanities
- 3:15-4:00: Session 3: Going further with GIS

These times could be extended a little if required to allow us to explore issues in a little more depth.

Content:

Session 1: Fundamentals of GIS from a humanities perspective

This session consists of two talks of around 20 minutes each leaving an additional 20 minutes for questions. Both presentations will be given by Gregory. The first of these talks will define what GIS is taking a technical perspective, introducing the core terminology and the data models that allow GIS to model the world. It will: define GIS, introduce concepts such as georeferencing, layers, raster and vector data, spatial and attribute data, querying GIS data, and the use of maps within GIS. The second talk will take a top-down approach giving examples of what GIS has to offer to the humanities. It will argue that the four benefits that GIS gives to the humanities are the abilities to: structure, integrate, visualise and analyse spatially referenced data. A wide range of brief examples will be used to describe what this offers in practice.

Session 2: Case studies on the use of GIS in the humanities

The second session will draw on two case studies on different uses of GIS within the humanities. The first will be presented by Dr Elton Barker a classicist from the Open University. Dr Barker will talk about his work on Mapping the Ancient World focusing on the work and travels of Heroditus, a scholar from Ancient Greece. The second will be presented by Dr Anouk Lang, a literary scholar from the University of Strathclyde. Dr Lang has used GIS in a range of literary studies focusing on modernist writing outside Britain and the United States. As well as being from contrasting disciplines and taking different approaches to using GIS, these two scholars have approached using GIS in two very different ways: Barker's work has been collaborative, while Lang's has predominantly followed the lone scholar model. In this way we will demonstrate what can reasonably be achieved by academics using GIS in the digital humanities. Again, presentations will be around 20 minutes giving plenty of time for discussions about the issues raised.

Session 3: Going further with GIS

Session 3 is a slightly shorter session. The intention is to point researchers to where they need to go next if they want to get more involved in GIS and to allow them to raise the issues and questions that concern them. It will consist of a short presentation by Gregory that will briefly introduce issues associated with software and access to data. It will then broaden out into a round table discussion to allow the participants to ask the questions that they want to ask. Past experience has shown this to be one of the most valuable parts of the day.

Format:

The workshop is best done as a half-day. It can be accommodated in three hours if required, however four would allow a more in-depth approach. There will be no hands-on software practicals so all that is required is a room where speakers can present to an audience. We want to encourage discussion among the participants so a tiered lecture theatre would not really be suitable.

Target audience:

The workshop is aimed at a broad audience including post-graduate students, members of academic staff, curriculum and research managers, and holders of major grants and those intending to apply for them. The workshop is only intended as an introduction to GIS, so will suit novices or those who want to brush up previous experience. Based on past experience we would anticipate that the workshop will be popular and would suggest capping numbers at 30.

Communication and publicity:

No CFP will be required and the organisers are happy to publicise it beyond the conventional DH mailing lists. When running similar events in the past we have asked participants to include a paragraph in their application about why they want to come to the workshop. This has two advantages: it allows us to tailor the content more closely to their needs, and if we are over-subscribed, it allows us to choose which participants will benefit most from the workshop. If this is possible then we would like to do this again, if not we can use a first come, first served approach.

The instructors:

Prof Ian Gregory is Professor of Digital Humanities at Lancaster University. His research concentrates on the use of GIS in the humanities. He is currently Principal Investigator on the ERC-funded *Spatial Humanities: Texts, GIS, Places* project. He has published widely on the theory and applied use of GIS in the humanities including *Historical GIS: Techniques, methodology and scholarship* (CUP, 2007); *Troubled Geographies: A spatial history of religion and society in Ireland* (IUP, 2013) and *Towards Spatial Humanities: Historical GIS and Spatial History* (IUP, in press).

Contact: Department of History, Lancaster University, Lancaster, LA1 4YT, UK. Tel: +44 (0)1524 594967. Email: I.Gregory@lancaster.ac.uk

Dr Elton Barker is a Reader in Classical Studies at the Open University. His research interests include: the Agon in ancient Greek literature and thought; Greek epic rivalry and reception; ancient geographies; and digital Classics. He has been Principal Investigator of three projects: the *HESTIA* project, *Google Ancient Places and Pelagios: Enable Linked Ancient Geodata*. He has published widely including *Entering the Agon* (OUP, 2009)

Contact: Faculty of Arts, The Open University, Walton Hall, Milton Keynes, MK7 6AA, UK. Tel: +44 (0)1908 653247. Email: Elton.Barker@open.ac.uk

Dr Anouk Lang is currently a Lecturer in English at the University of Strathclyde; from September 2015 she will be a Lecturer in Digital Humanities at the University of Edinburgh. She uses digital humanities and geo-spatial technologies to research modernism in the Anglophone world beyond Britain and the United States, and also to investigate contemporary reading cultures in digital contexts. She is the editor of *From Codex to Hypertext: Reading at the Turn of the Twenty-First Century* (U Massachusetts P, 2012), and is currently Principal Investigator for an AHRC-funded project on the cultural value of literary participation in digital

Contact: School of Humanities, Lord Hope Building, 141 St James Road, Glasgow G4 0LT, UK. Tel: +44 (0)141 548 3518. Email: Anouk.Lang@strath.ac.uk

Hettel, Jacqueline

Stanford University Libraries, United States of America

Lindblad, Purdom

The British Library, United Kingdom

Baker, James

University of Virginia, United States of America

Stack, Padraig

National University of Ireland–Maynooth, Ireland

Gil, Alex

Columbia University, United States of America

Miller, Laura

University of Virginia, United States of America

Bourg, Chris

Stanford University Libraries, United States of America

Description

This past summer at DH 2013 in Lincoln, Nebraska, we were all reminded that digital humanities and libraries have quite the longstanding relationship when an ADHO Special Interest Group was dedicated to Digital Humanities and Libraries. This relationship is especially profound as the library community is becoming increasingly more aware and concerned about creating opportunities for librarians to develop more skills and opportunities for training in the technical, prestige-laden skillsets. One of the goals of this group is to encourage and develop training projects in digital scholarship for library staff. Digital Humanities, and more generally digital scholarship, re-skilling and training opportunities provide librarians hands-on experiences with the skills and tools needed to create digital humanities projects, their existing skills are enriched, they are challenged to become fluent in less familiar areas, and most importantly they are empowered to become true partners in this field of scholarship—rather than simply purveyors of content. Furthermore, these efforts of growing sustainable DH initiatives within libraries help to ensure the long-term survival of digital humanities.

This one-day workshop is an opportunity for digital humanists working with/in libraries to come together to learn from one another about current international efforts for training library staff to acquire digital skills relevant to emerging digital research and scholarship practice. The goal of this workshop is that participants will be able to systematically design an approach that works specifically for their home libraries, propose their approach successfully to library leadership, and come together as a community of practice to support future re-skilling initiatives.

In order to achieve these objectives, this workshop will begin with a panel-led discussion of five, international DH training initiatives for library staff. Although each of these initiatives have different foci and implementation methods, librarians with diverse backgrounds and the critical skills necessary to conceptualize, design, and implement digital initiatives are now vital members of teams able to create lasting digital projects both for library and scholarly use. Exposure to first-hand accounts of the development and implementation of these types of initiatives is vital before proceeding with initiative design because it will challenge participants' notions of re-skilling within libraries and possibly inspire ideas for how to implement such strategies within their own organizations.

The second half of the workshop will focus on leveraging the experiences and best practices shared by the morning panel in order to help each participant create their own strategic plan for a DH training initiative within their own library. This will be performed through an interactive exercise where participants will be walked through the design thinking process and then asked to work together to troubleshoot one another's plans for clarity and execution. We will then come back together as a group to discuss how to take these plans and successfully pitch them to library leadership for implementation. The day will be concluded with a discussion of how the day's efforts can be made more successful with the support of community:

Methods for Empowering Library Staff through Digital Humanities Skills

specifically the creation or integration of these many different and varied library staff education initiatives, that exist within specific institutional contexts, into a singular, sustainable community of practice for supporting efforts for empowering library staff to become more active collaborators in digital humanities.

Target Audience

The target audience consists of digital humanists working with/in libraries who are interested in developing opportunities for re-skilling and training in digital scholarship for library staff.

Participants are asked to prepare the following materials with them to the workshop:

1. A chart or map of your library's organizational structure;
2. A list of the needs of your library regarding digital humanities skills. This could be inspired by upcoming projects, recent initiatives outlined by library leadership, etc.
3. A list of five people within your own library who are experts in five different areas of digital scholarship. Participants are encouraged to think outside of the box regarding this. For example, is there someone in your metadata department that has experience in leveraging metadata standards? Do you have a phenomenal project manager working on digital library projects?

Workshop Leader Bios

James Baker

Curator, Digital Research
The British Library

James Baker is a Curator in the Digital Research team at the British Library, a group whose responsibilities include the delivery and development of the library's ongoing Digital Scholarship Training Programme. He has a Ph.D. in History from the University of Kent, where he is now an Honorary Research Fellow of Digital History, and retains an active research profile in the field of eighteenth century British comic art. Prior to his current position James held a Postdoctoral Fellowship with the Paul Mellon Centre for Studies in British Art, was a Project Manager for the City and Region project, and worked on Open Access advocacy at the Templeman Library, University of Kent.

Chris Bourg

Assistant University Librarian for Public Services
Stanford University

Chris Bourg is the Assistant University Librarian for Public Services at the Stanford University Libraries. Before coming to Stanford, Chris spent 10 years as an active-duty US Army officer, including 3 years as a member of the faculty at the United States Military Academy, West Point, NY. She is a graduate of Duke University, and holds an MA and PhD in Sociology from Stanford University. She has published on issues related to diversity and leadership, and shares her thoughts on academic libraries and higher education on her blog, Feral Librarian and on twitter as @mchris4duke.

Alex Gil

Digital Scholarship Coordinator
Columbia University

Alex Gil is Digital Scholarship Coordinator for the Humanities and History at Columbia. He serves as a consultant to faculty, students and the library on the impact of technology on humanities research, pedagogy and scholarly communications. Current projects include an open repository of syllabi for curricular research, an aggregator for digital humanities projects worldwide and other initiatives at the intersection of technology and the humanities. He is currently vice-chair of global-outlook::digital-humanities (GO::DH) and the organizer of the THATCamp Caribe series. His scholarly heart remains betrothed to Caribbean Literature in the 20th Century.

Jacqueline Hettel

Textual Research Librarian
Stanford University Libraries

Jacqueline Hettel is the Textual Research Librarian at Stanford University Libraries. She received her Ph.D. in

English from the University of Georgia, and during her time there was the Chief Research Assistant for the Linguistic Atlas Projects. Many of the projects she currently works on involve the development of sustainable webbased tools and resources for legacy humanities data, as well as sustainable strategies for empowering library staff and patrons to become better collaborators for digital scholarship and research. She is also an active researcher in the field of corpus linguistics and is interested in its application to domains outside of the classroom: i.e. the academic research library or the corporate boardroom. She blogs about many of these things at www.linguabrarian.com and on twitter as @jacquehettel.

Purdom Lindblad

Head of Graduate Programs
Scholars' Lab, University of Virginia

Purdom Lindblad is the Head of Graduate Programs with the Scholars' Lab at the University of Virginia, where she coordinates the Graduate Fellowship in Digital Humanities and the Praxis Program. She is involved with broader graduate initiatives, including the Praxis Network. She has a MA in American Studies from Michigan State University and a MSI from the University of Michigan. Throughout her graduate education, Purdom worked with Matrix: center for digital humanities and social sciences at Michigan State University. The rich combination of disciplinary inquiry, information science, and digital humanities shaped her interest in emerging methodologies and digital practices for humanities graduate education. She is on twitter as @Purdom_L

Laura Miller

Digital Scholarship Services Librarian
Scholars' Lab, University of Virginia

Laura Miller is the Digital Scholarship Services Librarian in the Scholars' Lab at the University of Virginia. In her role as head of public programs for the Scholars' Lab, she is actively involved in an ongoing initiative to better leverage librarians as expert partners in the research process. With a background in literature and a MLIS from Florida State University, her interests include data management, changing models of scholarly publication, and user-centered design.

Padraic Stack

Digital Humanities Support Officer
NUI Maynooth

Padraic Stack is the Digital Humanities Support Officer for the National University of Ireland, Maynooth. His role is split between the University Library and An Foras Feasa, the Institute for research in Irish Historical and Cultural Traditions. He has an M.A. in Digital Humanities and is an Associate of the Library Association of Ireland. Prior to his current position he has worked in libraries in the private, public and community sectors. He is interested in the representation of history online, in the enrichment of archives through other sources and in the collection and distribution of informal histories.

Workshop Schedule (6 hours--full day)

Morning Session (3 hours) --

Why DH re-skilling/training initiatives are important for libraries. (Jacqueline Hettel)

9:00-9:15	Discussion around this topic and brainstorm as a group what challenges we face in developing re-skilling/training initiatives.
-----------	--

Different approaches to these initiatives: (2 hours)

9:15-9:20	Introduction of Panel/Projects
9:20-9:40	Instructional--Digital Scholarship Training Programme at the British Library (James Baker)
9:40-10:00	Experiential--Developing Librarian at Columbia (Alex Gil)
10:00-10:20	Experiential--DH Training for Library Staff at NUI Maynooth (Padraig Stack)
10:20-10:35	Break
10:40-11:00	Hybrid (Instructional/ Experiential)--Library Praxis at UVa (Laura Miller) Hybrid (Instructional/Experiential)--<digiPrep> at Stanford (Jacqueline Hettel)
11:30-12:00	Q&A with the Panel About Their Experiences Developing Training Initiatives
12:00-14:00	Lunch
14:00-15:30	Training Initiative Design Workshop--Creating a Tactical Strategy Leveraging Strengths & Resources Already in Your Library (Jacqueline Hettel)
15:30-15:45	Break
15:50-16:20	Onboarding Library Leadership in Supporting Reskilling Initiatives (Chris Bourg)
16:30-16:50	Creating a Community to Support These Ongoing Efforts (Purdom Lindblad)
16:50-17:00	Closing Thoughts/Thanks

Introduction to Text Analysis and Topic Modeling with R

Jockers, Matthew

University of Nebraska-Lincoln, United States of America

General Description:

Introduction to Text Analysis and Topic Modeling with R is a sequence of two workshops that will provide a practical introduction to text analysis with a special emphasis on topic modeling. Taken together, the workshops will cover basic text processing, data ingestion, data preparation, and topic modeling. The main computing environment for the workshops will be R: "the open source programming language and software environment for statistical computing and graphics." While no programming experience is required, students must have basic computer skills, must be familiar with their computer's file system, and must be comfortable entering commands in a command line environment. Though the two workshops are designed to stand alone, the second one is more advanced and assumes some basic familiarity with topic modeling. Participants might want to visit The LDA Buffet is Now Open; or, Latent Dirichlet Allocation for English Majors for a general overview.

Suggested Workshop Preparation:

While not required, participants are encouraged to work through at least the first two of the seven basic R lessons available at R Code School prior to taking this workshop.

In advance of the workshop, students should:

1. Download the current version of R (at the time of this writing version 3.0.0) from the CRAN website by clicking on the link that is appropriate to your operating system (see cran.at.r-project.org):
 - If you use MS Windows, click on the "base" and then on the link to the executable (i.e. ".exe") setup file.
 - If you are running Mac OSX, choose the link to the most current package.
 - If you use Linux, choose your distribution and then the installer file. Follow the instructions for installing R on your system in the standard or "default" directory. You will now have the base installation of R on your system.
 - If you are on a Windows or Macintosh computer, you will find the R application in the directory on your system where Programs (Windows) Applications (Macintosh) are stored. If you want to launch the R GUI, you can double click the icon to start the R GUI. We will not be using the R GUI in the workshop. We will use RStudio (see below).
 - If you are on a Linux/Unix system, simply type "R" at the command line to enter the R program environment.
2. Download and Install RStudio
 - The R GUI application is fine for a lot of simple programming, but RStudio is an application that offers a very nice user environment for writing and running R programs. RStudio is an IDE, that's "Integrated Development Environment" for R. RStudio runs happily on Windows, Mac, and Linux. After you have downloaded R (by following the instructions above) you must download the "Desktop" version (i.e. not the Server version) of RStudio from www.rstudio.com . Follow the installation instructions and then launch RStudio just like you would any other program/application. When you launch RStudio, you do not have to also launch the R program. RStudio accesses the R program you installed in the first step.

Workshop Syllabus:

Important: It is critical that you arrive on time to every session and be ready to roll with RStudio installed and running. The workshop will begin on schedule, and if you miss the first few minutes of any session you'll be lost!

Workshop One: Introduction to Text Analysis with Applications in R

Summary: In this workshop you will be introduced to the R programming language while learning the basics of computational text analysis. You will learn basic R syntax and be introduced to the RStudio programming environment. Text analysis topics covered will include text ingestion and tokenization, word frequency analysis, dispersion plots, and if time permits, correlation analysis.

- *Session one* (9:00-10:15)*
 - The R computing environment
 - R console vs. RStudio
 - Basic text manipulation in R
 - Word Frequency
- *Break* (10:15-10:30)
- *Session two* (10:30-12:00)
 - Dispersion Plots
 - Correlation

* Times are for illustration only.

Workshop Two: Introduction to Topic Modeling with Applications in R

Summary: In this workshop, you will be introduced to topic modeling and learn how to analyze and visualize topic model output in R. For this work, we will use the R implementation

of MALLET that was developed by David Mimno. Student will also learn how to parse TEI-based XML and how to segment large texts into chunks. We will discuss various text pre-processing procedures including how to do part of speech tagging in R using the openNLP package. Though this will be a hands-on workshop, some techniques explored here are quite advanced and those unfamiliar with such things as XML document structure and basic text analysis may find it better to observe and then use the included documentation to practice the techniques at home.

- *Session three* (13:00-14:30)*
 - Loading a corpus
 - Preparing files for Topic Modeling
- *Break* (14:30-14:45)
- *Session four* (14:45-16:00)
 - Running the Model
 - Exploring topic coherence with term clouds
 - Topic data analysis

* Times are for illustration only.

Project management and sustainable revenue models in the Digital Humanities

Keller, Stefan Andreas

stefan.andreas.keller@phil.uzh.ch
Faculty of Arts, University of Zurich

Keller, Alice

Zentralbibliothek Zurich

Neuroth, Heike

Niedersächsische Staats- und Universitätsbibliothek Göttingen DE

Rosenthaler, Lukas

DHLab, University of Basel

1) Topic & Motivation

The Workshop is the european „successor“ of the last-year „Sustainability Strategies“- Workshop in Nebraska at DH2013 ran by Nancy Maron from Ithaka S+R (<http://dh2013.unl.edu/schedule-and-events/workshops/#sustainability>). The workshop@DH2014 offers a training in projects management and sustainability like the Ithaka-Workshop, additionally combined with a Round Table joined by national and international key players (Humanists, librarians, archivists, members of funding organizations etc.) to discuss main problem areas and to find possible solutions. The Project is supported by several key institutions in the Digital Humanities like the US-Office for Digital Humanities (ODH) [Brett Bobley] or the King's College in London (Tobias Blanke).

Projects in the Digital Humanities are facing a lot of challenges concerning project management and sustainability in contrast to „traditional“ academic researchs projects. The limits between research, teaching and what one may call a lasting online service are blurred; projects have to deal with the laws of the field of the fast moving, fluid and competitive web-business. But because projects in the Digital Humanities are founded by academics, they are shaped by a research-centered and a supply-side mentality, which is neither aimed for sustainability nor user-oriented. Many of these projects — let's call them generally „Online Academic Resources“ — state problems developing a specific and adequate project management as well as a sustainable revenue model. Project leaders are simply not trained for such an enterprise. Coupled with the fast pace of the WWW and/or its technology a partial or complete absence of appropriate project management is to be stated. Project leaders seem to be overtaxed and thus the projects fail or cannot bail out their often huge potential. Furthermore, especially in switzerland, there is lacking an

openness to collaborate with qualified partners, e.g. librarians and archivists who could bring in a lot of skills and experiences.

The Workshop aims at helping project leaders to improve project management, to develop sustainable revenue models and to bring different people together to share Knowledge and Know-How. Core goals of the Workshop are therefore: Offer a training possibility for project leaders and other interested people, connecting people and sharing Know-How, create an awareness of the relevance of the topic with key partners in Switzerland – because here, unfortunately, currently there is no organization yet that specifically supports Digital Humanities projects – in contrast to the EU and U.S. space. Of particular relevancy would be the participation of libraries and archives, since they are playing an increasingly more important role in the world of Digital Humanities.

B) Structure: Tutorial & Round Table

The Workshop will be divided into two parts of approximately two hours each: 1) A training tutorial in project management and sustainability ran by Stefan Andreas Keller with a maximum of 20 participants (recommended maximum number of participants is based on the experiences of the Ithaka-Tutorial). Participants of this tutorial should be those with interest in and/or responsibility for charting a course for the development of a digital scholarly project or resource. This could include:

- Academic project leaders who are leading or have created a digital resource.
- Managers of digital collections and digitization units at cultural organizations, including museums, libraries, archives and other institutions.

Those in early stages of considering sustainability strategies for their projects are encouraged to attend.

Some participants like important project leaders (especially from Switzerland) will be invited prior to the Workshop, but it is intended to leave some places open for participants of the DH-Conference to join - also spontaneously - the workshop on site.

2) A Round Table with invited national and international key partners to present strategies and discuss solutions on the topic. Experts from different countries (DARIAH, ODH, Ithaka S+R) will give small inputs concerning the specific situation in their countries and about infrastructures where project leaders can get support and funding. In addition, there should be an open discussion on different relevant topics between project leaders, DH-institutions and potential sponsors. Several relevant topics could be discussed and fixed in advance or directly on-site. Number of participants will be at about 20.

Some of the central questions of the workshop are:

- Which areas of project management can be isolated in consideration of the different types of Digital Humanities-Projects?
- What models of management and revenue do already exist and in which way are they adaptable for others?
- What are the (dis-)similarities between projects? Is it possible to derive conclusions for project management and financing at all?

Concerning the legal and organizational level:

- What are the roles and the relationships of/between humanists, editors and libraries?
- Which type of project is most suited to be ran by one of these roles ?
- What are the legal frameworks (Copyright, licensing etc.)?

C) Participants and Contributors (short-term list, to be completed):

1. Dr. Stefan Andreas Keller, Faculty of Arts, University of Zurich (CH)
2. Dr. Alice Keller, Zentralbibliothek Zürich (CH)
3. Dr. Heike Neuroth, Niedersächsische Staats- und Universitätsbibliothek Göttingen (DE)
4. Prof. Dr. Jan Christof-Meister, Universität Hamburg, (DE)
5. PD Dr. Lukas Rosenthaler, DHLab Universität Basel
6. Dr. Tobias Blanke, King's College London (UK)

7. N.N., Office for Digital Humanities, National Endowment for the Humanities (USA)
8. Prof. Dr. Johannes Stiegler, Universität Graz (A)
9. Dr. Sacha Zala, Diplomatische Dokumente der Schweiz (Dodis.ch) [CH]
10. Dr. Christophe Koller, IDHEAP - BADAC (www.badac.ch) [CH]
11. Daniela Vaj, Viaticalpes (www.unil.ch/viaticalpes) [CH]
12. PD Dr. Christophe Flüeler, e-Codices (www.e-codices.unifr.ch) [CH]
13. N.N., Schweizerischer Nationalfonds (CH)
14. N.N., Staatssekretariat für Bildung und Forschung SBFI (CH)

Linked Data and Literature: Encoding the Facts in Fiction

Lawrence, Katharine Faith

King's College London, United Kingdom

Summary

Is it possible to tell a story to a computer so that it can process the events that have happened? Can we computationally differentiate the Scotland of Tam Lin, Macbeth, Harry Potter and Brave while still acknowledging the shared concept of 'Scotland'? Or deal with Watson's wound being in both his leg and his shoulder in Conan Doyle's Sherlock Holmes? This workshop will provide a theoretical and practical introduction to the modelling of narrative elements for computational processing and look at the advantages and limitations of annotating stories in this way. Drawing on structuralist theory, philosophy, computer science and media studies this workshop is aimed at researchers working with narratives, especially fictional (or debateably fictional) narratives, and who are interested in how linked data techniques could open new possibilities for analysis and distant reading.

Workshop Description:

Prof. Hendler, one of the luminaries of the semantic web/linked data movement, illustrated the potential power of the semantic web by posing the following question as an example of a query that the technology should be able to answer: "what was that movie with the short henchman who decapitates a statue with his bowler hat?". This type of query is comparatively simple for a human to understand many will immediately be able to name the film (and more will be able to narrow it down as a Bond movie), but from the perspective of computational analysis the question is very complex. Students of comparative literature or mythology may dream of being able to search, if not for short men in dangerous bowler hats, then for stories where the world is created from someone's body parts (list by body part) or the moment when a spell is broken by a kiss. This workshop will enable attendees to take the first steps towards creating systems which will allow for this type of semantic querying.

From narrative structuralists to TVTropes, from the Bechdel-Wallace Test presenting a thinking point on representation to the role of specific events such as transformations, social situations or a climactic kiss stories give us a series of moments which are of interest both to researchers and to the wider world. The addition of computational techniques to the study of narratives has resulted in a breadth of distant reading which would have previously been beyond the realms of a single researcher. The linguistic analysis of the text is now relatively commonplace with easy to use tools available to extract statistical information about the composition. More advanced techniques bring the power of natural language processing, annotation, word stemming and synonyms into play

to allow researchers the opportunity to reveal the structure of the telling of the story more efficiently than ever before. This collected and processed data can also drive subject indexes, making digital texts more accessible to the researcher than ever before. However the limitations of such techniques is that they work at a surface level and barely brush the semantics or structure of the story encoding of the story elements as linked data is one way to address this issue.

This workshop will focus on fictional narratives because they present numerous challenges beyond those shared with non-fictional narratives including the malleable nature of reality and how we can deal with the idea of truth within fiction. Since modelling is often seen as problematic because it is reductive in nature, the workshop will address the role of the model and the tension between the requirement to formalise to make the data computational and the inaccuracy and loss of information that is inherent in that process.

We will discuss the effect that levels of granularity and expectation have on model use in narrative study and the concept of the computer as an unreliable narrator.

Using examples and hands-on activities, this workshop will take attendees through the steps needed to extract and define story elements in a meaningful semantic way. While the exercises will focus on the OntoMedia ontology, other ontologies such as the Proppian Fairytale Markup Language (PftML) will be introduced and attendees who have already begun work in this area will be encouraged to share their experiences and models. While the examples will be text based, the standoff nature of linked data allows us to apply the same techniques to fiction in many forms of media and, indeed, across multiple medium. Stories have existed and do exist in every format that humans have created from the earliest oral tradition to the latest Hollywood blockbuster, and everything in between. They also do not exist in a vacuum. Intertextuality is an important part of narrative so being able to link between stories expressed through different media is vital. The workshop will give attendees the opportunity to consider ways of dealing with, and linking between, story and character variants.

Intangible culture, such as that represented in fiction, is increasingly recognised as an important part of our heritage. In promoting ways for researchers to record, publish, share and analyse this data in open ways, this workshop encapsulates the conference theme of digital cultural empowerment. It also asks us to think about the way in which we classify media content and how the push to identify and filter by content may have unexpected repercussions on how, when and why we annotate narratives.

Attendees will be asked to bring laptops for use during the practical sessions but will have the option of working in pairs. Short texts will be provided for use in the practical sessions and it is not expected that attendees will need to install any software prior to the workshop. While the workshop will work with linked data and ontological models, the focus of the modelling discussion will be on the theoretical side and knowledge of OWL and other modelling languages will not be required. Some familiarity with XML and RDF may be helpful as attendees may be asked to work with source code but the workshop is intended as an introduction and no knowledge/ability beyond basic computer use will be assumed.

Presenter:

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

Dr Lawrence is a Research Associate at the Department of Digital Humanities, King's College London where she works as a researcher and developer on a number of projects. Her research background centred around online communities, narrative and the semantic web. Her thesis, '*The Web of Community Trust Amateur Fiction Online: A Case Study in Community Focused Design for the Semantic Web*', investigated usercentred design for emergent technologies through the case study of online fiction archives and author communities. This work focused on fan fiction communities, both in terms of how they currently interact with technology, and how that interaction may evolve in the future with the development of Web 2.0 and the semantic web. One important facet of this work was an investigation into the description

of narrative and content elements within textual, visual, aural and multimedia works. She is one of the cofounders of the OntoMedia ontology for describing narrative in heterogeneous media.

Target Audience: narratologists, literary scholars, historians, folklorists, media scholars, oral historians

Expected number of participants: 20 - 30

Outline of Content:

Morning:

- Welcome and Introduction (30 mins)
 - Presentation: The Good, the Bad and the Ugly of Modelling Narrative Elements in Fiction (45 mins)
 - Group Activity: Annotating Little Red Riding Hood I - Identifying the Narrative Elements (30 mins)
- [Break]
- Report back and Discussion (30 mins)
 - Presentation: And Then Something Happened Granularity and Defining Events (45 mins) Afternoon:
 - Group Discussion: Extracting Events Examples (30 mins)
 - Group Activity: Little Red Riding Hood II Typing the Elements (30 mins)
 - Presentation: Can we handle the truth? Variation, Intertextuality and Unreliable Narrative (45 mins)

[Break]

- Group Activity: Little Red Riding Hood III Linking the Elements (30 mins)
- Presentation: Exploring the data (30 mins)
- Wrap up (15 mins)

Length: 1 Day

Linked data has become an increasingly popular fixture in digital humanities research because it offers a way to break out of the data silos that are constantly being created, and provides a framework for new ways of approaching research questions. Tim BernersLee's four principles of linked data, however, remind us that global identifiers for entities URLs – provide only a part of what is needed if linked data is to fulfill its promise. As much as possible, we also need common semantic frameworks to better tie the data together – what are called "ontologies".

In a seminal paper way back in 1993 Thomas Gruber defined an ontology as an "explicit specification of a shared conceptualisation". We will be focusing on possibilities for an ontology for prosopography because, for historical data at least, people, places and textual sources are likely to be the three pillars upon which a structure of linked data can be constructed, and these three things are likely to be the primary entry point for a collection of linked historical data. While methodologies for dealing with textual sources are being continually refined, the success of the Pelagios project has demonstrated how historical geographic information, in this case classical, can be used to bring together a wide variety of projects. This workshop will address the issues of bringing linked data to the description of historical persons with the morning session devoted to exploring the question of whether there are sufficient common concepts – a shared conceptualisation – to enable for the practical and useful development of an ontology for historical persons, and the afternoon addressing the challenges of linking these descriptions together to create a shared resource.

In the morning we will be following up on Gruber's recognition that the best way to define an ontology is to look for shared conceptualisations by examining the practices of a range of existing, or emerging, projects that attempt to capture information about historical persons using structured models that are compatible with semantic web thinking. We will present a detailed introduction to a number of the significant models currently in use including the data model behind the University of Virginia's *People of the Founding Era*, the factoid model used for a number of prosopographical projects from King's College London, the SNAP:DRGN relationship model, the prosopographical components in the well-known CIDOC-CRM and FRBRoo, and will explore the developing standards for archive data, starting with University of Virginia's *Social Network and Archival Context* (SNAC) model (and its prototype site), to the standards emerging from the International Council on Archives Experts Group on Archival Description. Additionally, workshop participants will be encouraged to share any models they have used for digital prosopography, and their views about the models we present. This session aims to give those attendees who are new to question of linked data and prosopography an introduction to the subject while offering the opportunity for those with existing data to discuss and compare the approaches with a view towards identifying best practice and whether a standard model for describing historical persons is possible.

The afternoon portion of the workshop will focus on the publication and linkage of prosopographical data. The Quantified Authenticated CoReference (QuAC) data model being developed by the *Standards for Networking Ancient Prosopography* (SNAP) project for the sharing and linking of names, persons and person-like entities in historical data. The SNAP model is being tested with existing digital resources, including *Prosopographia Imperii Romani*, the *Lexicon of Greek Personal Names*, and *Trismegistos People*, and working with a wide range of other projects. One of the key aims of SNAP is to model the complexity, uncertainty and ambiguity inherent in true prosopography, in contrast to the sometimes simplistic approaches of modern social media. The aim of this session is to allow more indepth, directed discussion and the opportunity for handson data hacking sessions through the use of breakout groups. Attendees will have the opportunity to work with technical facilitators to apply the SNAP model to their own or example data. For those who are more interested in the theoretical framework, facilitators will lead discussions building on the mornings activities and standards for modelling historical persons and on developing specifications for what services,

Ontologies for Prosopography: Who's Who? or, Who was Who?

Lawrence, Katharine Faith

King's College London, United Kingdom

Bodard, Gabriel

King's College London, United Kingdom

Bradley, John

King's College London, United Kingdom

Perdue, Susan

Virginia Foundation for the Humanities

Rahtz, Sebastian

University of Oxford

Daniel, Pitti

IATH (University of Virginia), University of Virginia. (SNAC, ICAEG)

Christian-Emil, Ore

University of Oslo

Summary:

Historical data about people, their names, their attributes, and their relationships is one of the most common types of data to expose and one for which is falling behind other areas in the move to the digital data publication and exchange. This workshop will address the issues of modelling historical persons with presentations and discussions on existing models towards finding whether a crossproject consensus on standards and best practice is possible. The workshop will continue with discussion on how we can link the person data from different projects together and will offer the opportunity for handson, breakout sessions for those who have data they wish to publish as linked data.

Workshop Description:

outcomes and requirements researchers would want in order to share and reuse historical person data.

This workshop is sponsored by two projects with different foci and covering very different historical periods:

People of the Founding Era (PFE), a Mellon-funded project at the University of Virginia, aims to apply a prosopographical approach to collecting and publishing the biographical content found in the correspondence of prominent and not-so-prominent individuals in the time of the founding of the United States. An important challenge in the project is identifying slaves who are not well represented in the documentary records. PFE is working with linked data as a means to establish identity and suggest connections between numerous anonymous or partially named people or for those who are known only by their occupation or owner.

SNAP.DRGN (Standards for Networking Ancient Prosopographies: Data and Relations in Greco-Roman Names), a project which aims to address the problem of linking together large collections of material (datasets) containing information about persons, names and person-like entities managed in heterogeneous systems and formats from the Ancient World.

What unites them is what unites many digital projects; the need to deal with historical data about people, their names, their attributes, and their relationships one of the most common types of data to expose and one for which is falling behind other areas in the move to the digital data publication and exchange. The collaboration between these two projects clearly demonstrates the importance of this subject to a wide range of digital humanities researchers and we believe that this workshop will encourage vital cross-disciplinary discussion about prosopography that emerges from different periods and cultures.

Speakers:

Dr Gabriel Bodard, King's Colledge London
(gabriel.bodard@kcl.ac.uk)

Bodard is the Principal Investigator of the SNAP:DRGN project. His research interests are in digital study, encoding and publication of classical texts, especially ancient Greek inscriptions. In 2004 he founded the Digital Classicist, a community of expertise in the application of Digital Humanities to the study of the ancient world, and is an administrator of the Stoa. He was on the steering committee of the British Epigraphy Society from 2007-2012, and was an elected member of the Technical Council of the TEI from 2008-2013, an academic group that makes decisions on guidelines and technical development. He is one of the lead authors of the EpiDoc Guidelines, and regularly organises and teaches training workshops in digital epigraphy and papyrology. He led the King's team on the internationally collaborative Integrating Digital Papyrology project (2007-2011) to convert the DDbDP and other papyrological materials into EpiDoc XML in a new browse and editing platform.

John Bradley, King's Colledge London
(john.bradley@kcl.ac.uk)

Bradley has for many years been involved in structured prosopography through seven prominent collaborative prosopographical projects including the Prosopography of AngloSaxon England (PASE) and the Peoples of Medieval Scotland (PoMS), and (although not its original inventor) has promoted the factoid model as a way to think about structuring prosopographical data. Recently he has taken up thinking about the place of prosopography in the context of global, open, linked data, and has given presentations on the idea at DH2013 and at the Culturecloud, Coreference and Archive Workshop given at the National Archives in Stockholm in June 2013.

Dr K Faith Lawrence, King's Colledge London
(faith.lawrence@kcl.ac.uk)

Lawrence is a Research Associate at the Department of Digital Humanities, King's College London where she works as a researcher and developer on a number of projects. Technical lead on the SNAP:DRGN project her research background centred around online communities, narrative and the semantic web. Her thesis, 'The Web of Community Trust Amateur Fiction Online: A Case Study in CommunityFocused Design for the

Semantic Web', investigated usercentred design for emergent technologies through the case study of online fiction archives and author communities. This work focused on fan fiction communities, both in terms of how they currently interact with technology, and how that interaction may evolve in the future with the development of Web 2.0 and the semantic web. One important facet of this work was an investigation into the description of narrative and content elements within textual, visual, aural and multimedia works.

Prof. Susan Perdue, Virginia Foundation for the Humanities (ssh8a@eservices.virginia.edu): (PFE)

Perdue is a documentary editor who has worked primarily in the American Early Republic. Her focus on name authority work began with print indexes and evolved to XML indexing and markup in historical documents. Begun in 2008, People of the Founding Era is a prosopographical project that aggregates content from hundreds of American Founding Era documentary volumes, supplemented with research. The project draws on the expertise of editors and museum professionals to centralize their longstanding research, especially that related to slavery in the Early Republic.

Bob DuCharme, TopQuadrant (bob@snee.com)

DuCharme is the author of Learning SPARQL from O'Reilly Media which introduces the reader to the W3C standard for querying Linked Data and the semantic web. He is also an expert on XML and XSLT and works at TopQuadrant, a leading software company in the semantic web world. For the past year, DuCharme has worked with Perdue and PFE to implement an RDF model that queries PFE data and other related data sources, called LDGES.

Sebastian Rahtz, University of Oxford

(sebastian.rahtz@it.ox.ac.uk) Rahtz is Director of Academic IT at University of Oxford University IT Services, where he oversees the teams responsible for research support and open source. He has been closely associated with the Text Encoding Initiative for the last decade as a member of its Technical Council, and architect of its metaschema system. Since 2008 he has been part of the team developing CLAROS ("the world of ancient art on the semantic web") at Oxford, for which he leads the Metamorphoses subproject to manage its place and name linking. He has worked with the Lexicon of Greek Personal Names at Oxford for the last 30 years, and maintains its experimental online service and data export.

Dr Daniel Pitti, University of Virginia (dpitti@virginia.edu)

Pitti is Associate Director of IATH (University of Virginia) and the chief technical architect of both the EAD and EAC-CPF standards, as well as being project director of the NEH and Mellon funded (Social Networks and Archival Context) SNAC project (2010-2015). SNAC is exploring the feasibility of extracting the descriptions of people that archivists routinely create when describing archival resources in order to maintain the descriptions independently though in relation to the records that are the evidence of the lives and work of the people described. As the chair of the International Council on Archives Experts Group on Archival Description, charged with developing a conceptual model for archival description, Pitti is also interested in how the descriptions of people created by archivists can be formalized and structured in such a manner that they can be shared with allied cultural heritage communities and scholars.

Dr Christian Emil Ore, University of Oslo

(c.e.s.ore@iln.uio.no)

Ore is an associate professor in the Department of Linguistics and Scandinavian Studies at the University of Oslo and is the head of their Unit for Digital Documentation. He has taken a keen interest in digital humanities for many years. He has been an active player in the CIDOC CRM community, one of the four current editors of the CIDOC CRM standard and has explored methods to combine TEI (Text Encoding Initiative) encoded documents with CIDOC CRM models.

Target Audience: Prosopographers, biographers, genealogists, classicists, social historians, (computer) ontologists, linked data/semantic web developers

Expected number of participants: 30 - 40

Outline of Content:

Morning: Modelling the Person

Welcome and Introduction (10 mins)
Group Activity Historical Speed Dating (30 mins)

Presentation and Discussion:

- The conception of prosopography in the PFE project, and its representation in RDF (20 mins)
 - A Semantic Web understanding of the factoid prosopography model (20 mins)
- [Break]
- Exploring prosopography in CIDOC CRM/FRBRoo, SNAC, and in the emerging standards from the International Council on Archives Experts Group on Archival Description (20 mins)
 - SNAP:DRGN: Going QuACers the Qualified, Authenticated Coreference model (20 mins)
- Round up and open discussion (1 hour)

Afternoon: Linking the Person

Welcome Back (15 mins)
Breakout 1 (1 hour):

- Breakout Group 1 SNAP Services: Discussion and User Requirements
- Breakout Group 2 Data Exchange and Chop Shop: Data Preparation Tutorial
- Breakout Group 3 Data Exchange and Chop Shop: Data Hacking
- Breakout Group 4 The historical person model

[Break]

Breakout 2 (1 hour):

- Breakout Group 1 - SNAP Services: Discussion and User Requirements
- Breakout Group 2 - Data Exchange and Chop Shop: Data Preparation Tutorial
- Breakout Group 3 - Data Exchange and Chop Shop: Data Hacking
- Breakout Group 4 - The historical person model

Reports and Discussion (30 mins)

Conclusion (15 mins)

Length: 1 Day

The Representation of Multiplicity as a Means to Digital Cultural Empowerment

Mareike, Hoeckendorff

University of Hamburg, Germany

Vitale, Valeria

King's College, London, United Kingdom

Dunn, Stuart

King's College, London, United Kingdom

Gius, Evelyn

University of Hamburg, Germany

This workshop is addressing the theme of DH 2014, "Digital Cultural Empowerment", by discussing the multiplicities inherent to the representation of cultural heritage and exploring possible ways of dealing with them.

The representation of cultural heritage needs to face multiplicity in many ways: it has to deal with different layers of time (from past to present), different types of objects on different levels of reality (physical "real" artifacts, digital representations of it, hypotheses about them, fictional

approaches etc.), and different views on it, depending on purpose, socialization, interests, knowledge etc. In many cases we even have to deal with things that existed in the past and now survive only in documentation, memory and/or imagination. On a more abstract level, then, the concept of multiplicity which we wish to explore is based on a definition of 'culture' as a dynamic experience.

On this backdrop empowerment means to represent and make accessible culture in a way that enables people to participate, to narrate their own story/stories about the represented, and to add their personal or professional interest and information. This is facilitated by the "empowerment" of cultural artifacts and practices: Pictures, stories, songs, movies and especially how people observe, describe and use artifacts, places etc. become equally valuable for the representation of culture.

Digital cultural empowerment thus means to represent cultural heritage in a multilayered way, enabling people to access different layers of raw data (images, literary texts, movies, music) as well as interpretations, mashups, comments and discussions. In a digital cultural environment people should both be enabled to passively explore what is exhibited and to actively use the data for their own narration about the represented—and even add more data and information.

Bringing the constructivist approach to a digital environment, the workshop thus aims to discuss projects that enable the users to actively participate, as single individuals as well as groups, to the process of the building of meaning, instead of passively consuming their own cultural heritage (Copeland 2004, Parry and Arbach 2009).

Our goal for the workshop is to tackle crucial issues in this context by engaging in an exchange with the participants. Presentations of approaches from various projects will boost the discussion of issues of multiplicity, representation, participation and information handling.

As a kickoff we will present two projects from the field and their approach to the addressed issues, afterwards we'll invite participants to present their approaches and to engage in discussions. The envisaged outcome of the workshop is a set of recommendations for digital cultural projects concerning the approach to issues of multiplicity, representation, participation and information handling described below.

The first exemplum project will be the *efoto* project, a cooperation between several cultural institutions in Hamburg and Hamburg University, brought into life by the Ministry for Cultural Affairs of the Free and Hanseatic City of Hamburg. The project deals with the urban space of the city of Hamburg represented by a huge supply of historic and contemporary photographs. *efoto* attempts to exemplify how the interactive use of digital media and services can enable the construction and discussion of individual and shared cultural experience, thus exploring and contributing to multiplicity at the same time.

efoto is based on Niklas Luhmann's conception of culture as a stock of themes that encourages communication and motivates interaction on a reflective level (Luhmann 1987, 1995). One of the main goals of *efoto* is thus to proceed from the level of the purely visual "cityscapes" to an exchange about what constitutes the culture of the city of Hamburg.

The core implementation component of *efoto1* is a digital database consisting of pictures of Hamburg, currently stored in numerous public and private archives, and metadata such as geoinformations. Image and text data, primary and metadata will be stored automatically and crosslinked so that they may be searched, combined and commented on according to user interests. This database will be enriched by several functionalities and services that enable users to participate in an open discourse on their urban living space. In order to achieve both—a high level professional database with a powerful intelligent search engine, and a multiinterest playground for leisurely use—three layers of data enrichment are anticipated. The first focuses on factual historical information, the second on personal experience based narratives and the third on the sociocultural practices of reflexion and interaction which in Luhmann's view is the essence of culture.

The second core component of *efoto* is a dedicated mobile application designed to support a wide range of use cases that contribute toward the main intention of the project i.e. getting

people to engage with cultural heritage in various ways, be they 'virtual' or 'real', 'professional' or 'private'. These multiple ways in which app users can appropriate and enrich primary image data about Hamburg are conceptually rooted in the practices of social annotation, social storytelling and in the crowd sourcing of primary data. Through their combination in *efoto* the multiple cultural identities of the city of Hamburg will take shape as a lived, dynamic experience of culture rather than as a static piece of cultural knowledge.

The second project deals with the representation of archaeological heritage, more specifically with the Temple of Isis, in Pompeii. Here the digital project aims to apply multiplicity to address two main issues:

The first is the hypothetical nature of all visualisations of archaeological objects. Ancient buildings, with their decorations and artefacts, can only be inferred by material clues and historical knowledge. Every virtual unification and restoration is, thus, highly speculative, even when based on the most rigorous research. However, when a very detailed (and often realistic) visual outcome is presented to the public, this is assumed to be the only possible and/or correct reconstruction, especially if it is endorsed by authoritative cultural institutions such as universities or museums. Presenting more than one visual hypotheses (developed by different researchers or even by the same one) highlights that all questions about the ancient world have many possible and legitimate answers.

The second and closely related issue is the interpretation of the Past. During the last 250 years, Pompeii has generated a vast amount of verbal and visual interpretations. This chronological overview, shows the interpretative process as a work in progress that is always influenced by cultural and social variables, and that has many commonalities with the practice of story telling.

Creating a digital tool that makes available, for a piece of archaeological heritage, both multiple restoration hypotheses and multiple interpretations, aims to promote a more critical approach to cultural heritage, in which, ideally, the public doesn't receive a single, simplified and unsatisfactory piece of knowledge from «the experts», but is directly engaged (and challenged) with the complexity of the topic and invited to experiment with different combinations and different readings of the source information.

After the presentation of these two exemplary projects we will invite participants to present and discuss their own approaches, experiences and/or findings related to one or more of the following issues addressed by both exemplum projects:

Multiplicity:

If we want to show the multiplicity every representation of culture has to deal with necessarily, we need to have answers to the following questions:

- How should multiplicity be addressed (generally or specifically)? Which are the most important aspects of it?
- Do we need to predefine different levels of quality of the represented according to its configuration regarding multiplicity issues (e.g. reality status)?
- How can we address the level of uncertainty connected to a certain cultural topic?
- What about contestability: Can/should multiplicity be used as a means of challenging one narrative by reference to another?
- Why does multiplicity matter?

Representation:

Every digital system dealing with culture should address the multiplicity inherent to the representation of culture by its design:

- Which tools are suitable for the visualization of which type of information?
- How can we represent technologically uncertainties, fuzziness or even contradictions stemming from multiplicity?

Participation:

Given we have a system that is capable of representing culture and its multiplicity in a conceptually and technologically adequate way:

- How can we engage people to participate in this digital cultural environment?
- What would participation mean then?
- (How) Can a digital cultural environment empower citizens?

Information handling:

Enhancing the data collected and prepared in a prevalently scientific context with user generated content means to add a new layer that is even more complex. Moreover, the concept of multiplicity developed won't be applicable to user generated content straight away:

- How can we deal with user generated content? Do we have to treat it differently?
- What tools are suitable for the extraction of information from the data, both provided and generated?

Workshop leaders:

Stuart Dunn, Centre for eResearch, Department of Digital Humanities, King's College London

Stuart's main research interest lies in Digital Geography and data visualisation, including theoretical aspects of Virtual Reality and agency theory. He is interested in how people, location and pace interact, and how those interactions can be expressed digitally. Another closely related field of interest is the perception, representation and interpretation of past environments, and how these can be reconstituted digitally, without imposing arbitrary constructs that are not, or cannot, be supported by empirical data.

Evelyn Gius, Department of Language, Literatures and Media, Faculty of the Humanities, University of Hamburg

Evelyn is currently working on the Digital Metropoles Network project, aimed at uniting projects from Florence, London and Hamburg that give digital access to culture. Before that she was involved in the development of the intermedial blended learning course "NarrNetz". She has been part of the development team of the computer markup tool CATMA for the last six years, too, and is engaged now in application of it in the heureCLÉA project which is focussed at the detection of narratological phenomena related to time.

Mareike Höckendorff, Department of Language, Literatures and Media, Faculty of the Humanities, University of Hamburg

is involved in the *efoto* project. Her research interest lies in cultural studies and contemporary literature and especially the relation between place, space and culture. In her PhD she analyses the representations of place in literature and how literature as cultural means takes part in the creation of atmospheric spaces in urban environments. Her case study is the city of Hamburg, her approach is to use digital tools and visualizations to transform a big amount of textual data into an easy to read interactive map showing the cultural landscape of the city in different stages of history.

Valeria Vitale, Department of Digital Humanities, King's College London

a digital humanist specialised in virtual cultural heritage. Her main research interest is the representation of the Past, and the methodological issues connected to the visualisation of something that does not exist anymore in its materiality. She's currently working on a communitybased ontology to document 3D visualisation for cultural heritage and display multiple reconstruction hypotheses for the same ancient object. In her PhD, she has recently started investigating the possible interactions between archaeological narratives and interpretations and 3D virtual environments. Her case study is the ancient city of Pompeii.

Workshop contact person:

Mareike Höckendorff
University of Hamburg

Department of Language, Literatures and Media
c/o Institute for German Literature and Language
VonMellePark 6
20146 Hamburg
+49 40 428382312
mareike.hoeckendorff@uni-hamburg.de

Target audience and expected number of participants:

The main audience of this workshop are scholars engaged in digital cultural projects and/or the development of technological means designed to tackle the issues described. In our experience a broad variety of participant experience and background enhances the quality of the discussions significantly, therefore we strongly encourage people from all project sizes and career stages to register for this workshop.

Special requirements for technical support:

There are no special requirements.

Length of the workshop:

one day

Provisional format:

- Introduction, overview of the workshop, outline of addressed issues
- presentation of approaches from the *efoto* project
- presentation of approaches from the Pompeij project
- discussion
- presentations by participants on projects (part 1)
lunchbreak
- presentations by participants on projects (part 2)
- presentations by participants on possible approaches
- discussion of the defined issues (in smaller groups or as plenary discussion, depending on interest)
- and number of participants)
- wrap up discussion

Program committee:

Jannis Androutsopoulos, Professor for German and Media Linguistics, Faculty of the Humanities, University of Hamburg
Stuart Dunn, Lecturer at the Centre for eResearch, Department of Digital Humanities, King's College London
Alexandra Georgakopoulou, Professor of Discourse Analysis and Sociolinguistics, Centre for Hellenic Studies, King's College London
Evelyn Gius, Research Assistant, Department of Language, Literatures and Media, Faculty of the Humanities, University of Hamburg
Mareike Höckendorff, Research Assistant, Department of Language, Literatures and Media, Faculty of the Humanities, University of Hamburg
Gertraud Koch, Professor for Folklore and Cultural Anthropology, Faculty of the Humanities, University of Hamburg
Markus Kuhn, Professor for Media Studies, Faculty of the Humanities, University of Hamburg
Jan Christoph Meister, Professor for German Literature, Department of Language, Literatures and Media, Faculty of the Humanities, University of Hamburg
Horst Scholz, Head of the Department for Information Technology of the Cultural Ministry of Hamburg, project leader of *efoto* at the Cultural Ministry of Hamburg
Caja Thimm, Professor for Media Studies and Intermediality, University of Bonn
Valeria Vitale, Research student in Digital Humanities, King's College London

References

- Luhmann, Niklas (1987). *Soziale Systeme: Grundriß einer allgemeinen Theorie*. Frankfurt am Main: Suhrkamp.
Luhmann, Niklas (1995). *Die Kunst der Gesellschaft*. Frankfurt am Main: Suhrkamp.
Parry, Ross and Arbach, Nadia (2006). *Localised, Personalised and Constructivist: A Space for OnLine Museum Learning*. In Kenderline, S. and Cameron, F. (eds). *Digital Cultural Heritage: a critical discourse*. Cambridge, MA: MIT Press.
Copeland, Tim (2004). *Presenting Archaeology to the Public*. In Merriman, N. (ed). *Public Archaeology*. London/New York: Routledge.
<http://www.efotohamburg.de/> and <http://www.jcmeister.de/projects/efotohamburg/> (last seen 201402.19).

Sound and (moving) images in focus – How to integrate audiovisual material in Digital Humanities research

Ordelman, Roeland

Netherlands Institute for Sound and Vision, Netherlands, The

Kemman, Max

Erasmus University Rotterdam, Netherlands, The

Kleppe, Martijn

Erasmus University Rotterdam, Netherlands, The

de Jong, Franciska

Erasmus University Rotterdam, Netherlands, The

Abstract

The proposed workshop intends to address the poor representation of audiovisual data in the evolving field of Digital Humanities. Sources such as television, film, photos and oral history recordings have not yet received the same level of attention from scholars as written sources. This can be considered as problematic in the light of the growth in volume of audiovisual sources in the near future, and the abundance of information that could be (re)used by various disciplines. In four sessions the workshop will discuss (a) issues related to the integration of audiovisual data in DH, (b) the necessary conditions and possible solutions, (c) examples of best practices and (d) an agenda for the future.

Introduction

Audiovisual material is perhaps the biggest wave of data to come in the near future (Smith, 2013). This claim is supported by a prospective study conducted by IBM on how the flow of digital data will evolve in the coming two decades. As can be seen in Figure 1 below, the development of audiovisual sources such as video, images and audio, will result in huge amounts of data in the coming decades, both due to the increased production of digital-born data and the massive digitisation of analogue sources. Consequently, audiovisual archives hold the promise of truly big data becoming available to academic researchers.

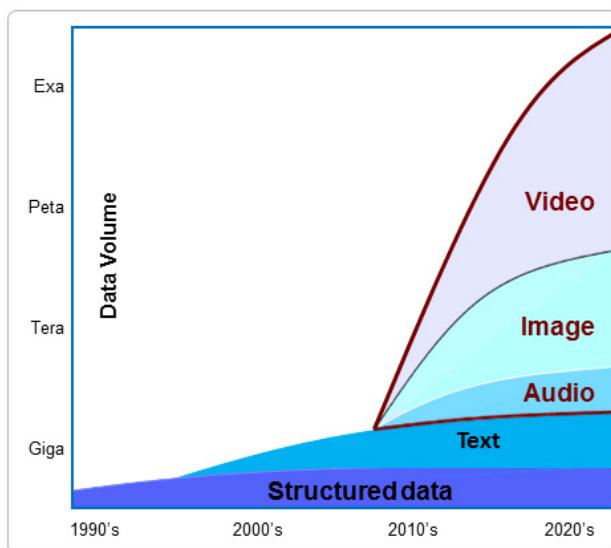


Fig. 1: Expected wave of data showing the growth of audiovisual data (video, images, audio) this workshop will deal with. Source: IBM Market Insights 2013

Audiovisual sources have a potentially huge value for the Digital Humanities as they are multi-layered. A single document can provide information regarding language, emotions, speech acts, narrative plots and references to people, places and events. This richness provides interesting data for various disciplines and holds the promise of multidisciplinary collaboration between e.g., computer sciences, social sciences and the humanities. As such, audiovisual material provides a rich playground for the Digital Humanities.

Notwithstanding this exponential growth, the use of audiovisual data by scholars in the social sciences and the humanities (SSH), and the application of digital methods for analysis are still in their infancy. Audiovisual material such as television, photos and oral history recordings have not yet received the same level of attention from scholars as written sources. Several reasons might account for this deficit. Firstly, the relatively young age of these source types compared to text; this is reflected in scepticism on their value for academic research outside a relatively small community of specialists. Secondly, the contemporary and commercial value of many audiovisual sources results in considerable constraints for use due to issues of copyright. Thirdly, the linear structure of audiovisual sources is problematic for hermeneutic analysis as it is more time-consuming compared to textual sources. Finally, no widespread accepted digital research methods for the discovery and analysis of audiovisual content exist as of yet. Unlike fellow scholars who study text and have a multitude of refined tools at their disposal, scholars specialised in documentaries, photo, film and audiovisual oral history collections, face considerable limitations in the various stages of the research process (De Jong et al., 2011). In the context of the proposed workshop, two themes will play a crucial role.

Theme 1: Indexing and searching audiovisual data

The first step in an SSH research process is the identification of relevant and interesting material. However, obtaining good search results is highly dependent on the richness and the level of granularity of the metadata assigned to the sources. Metadata is usually attributed to a document by a knowledgeable archivist. However, considering the sheer size of digital audiovisual content that is being produced daily, manual annotation is no longer feasible. Consequently, one of the first big challenges within the realm of audiovisual archives is the development of systems for accurate automatic annotation.

One could say that a revolution is needed similar to the one that full-text search (or automated text indexing) brought about. Content-based image retrieval has only recently made enough progress to be usable for scholars. Techniques such as speech

recognition and computer vision will support exploration of digital audiovisual archives on the basis of multiple modalities such as text, sound and image. However, this introduces the problem of the so called semantic gap, which refers to the difficulty of translating low-level pixel data and sound waves into meaningful annotations (Smeulders et al., 2000). How this semantic gap affects discovery of material in audiovisual archives is still under exploration.

Theme 2: Analysing audiovisual material

Besides identifying relevant content, an even bigger challenge on the side of the humanities and social sciences lies in providing tools for the next phase of the research process: the analysis and interpretation of content. While text mining has led to the phenomenon of distant reading of textual material (Moretti, 2013), which is strongly dependent on good visualisation tools, the advances in speech and image recognition have not yet led to a method of 'distant viewing' of audiovisual data. Processing large amounts of data and enabling researchers to trace patterns or discrepancies in their material are thus not yet feasible. Moreover, the lack of metadata which is often a feature of audiovisual archives introduces additional difficulties in heuristic practices (Fickers, 2012). Consequently, scholars working with (moving) images and sound are at a disadvantage in the evolving field of the Digital Humanities and effort has to be put in envisioning solutions.

Format

The proposed workshop aims to bring scholars and computer scientists together to discuss the following questions in four sessions.

1. Why are audiovisual archives scarcely used within the (Digital) Humanities? (Session 1)
2. What are possible technical solutions to stimulate the use of audiovisual archives within the (Digital) Humanities? (Session 2)
3. Which successful applications of DH on audiovisual data can serve as best practice? (Session 3)
4. Can we formulate a research and development agenda for a future uptake of audiovisual data in the (Digital) Humanities? (Session 4)

The keynotes within the first two sessions will be delivered by Prof. Andreas Fickers, who will talk about the use of audiovisual sources within humanities research, and Dr. Arjan van Hessen, who will discuss the necessary technical and infrastructural provisions for the analysis of these sources. For the third session we will invite scholars to submit papers and demos. The fourth workshop will focus on the evaluation of the findings and the formulation of an agenda for the future. To disseminate the results of the workshop among a broader audience, the initiators intend to propose a special issue on this topic to a Digital Humanities journal.

Acknowledgements

The proposed workshop is initiated by researchers working within the EU FP7 research project AXES – Access to audiovisual archives (www.axes-project.eu). We thank the AXES project for the financial support to organise the workshop.

Literature

Jong, F. de, Ordelman, R., & Scagliola, S. (2011). *Audio-visual Collections and the User Needs of Scholars in the Humanities: a Case for Co-Development*. In Proceedings of the 2nd Conference on Supporting Digital Humanities (SDH 2011). Copenhagen, Denmark.

Fickers, A. (2012). *Towards A New Digital Historicism? Doing History In The Age Of Abundance*. VIEW Journal of European Television History and Culture, 1(1), 19–26.

- Moretti, F.** (2013). *Distant Reading*. Verso Books.
- Smeulders, A. W., Worring, M., Santini, S., Gupta, A., & Jain, R.** (2000). *Content-based image retrieval at the end of the early years*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 22(12), 1349-138
- Smith, J. R.** (2013). *Riding the multimedia big data wave*. In Proceedings of the 36th international ACM SIGIR conference on research and development in information retrieval - SIGIR '13. New York, New York, USA: ACM Press. doi:10.1145/2484028.2494492

Digital Cultural Empowerment

Palm, Fredrik

fredrik.palm@humlab.umu.se
Umeå University, Sweden

Murphy, Orla

o.murphy@ucc.ie
University college Cork

Day, Shawn

shawn.day@ucc.ie
University college Cork

Thély, Nicholas

nicolas.thely@univ-rennes2.fr
Université de Rennes 2

Digital Humanities can act, and is acting, as an agent of digital cultural empowerment. Critical awareness of method and tools in Digital Humanities can bridge the gaps between the range of humanities corpora and the actual digital practices used to understand, analyse, and share them.

NeDIMA is about bridging these gaps and work groups on Space and Time, and on Information Visualisation, have successfully worked together across disciplines to consider these tools and methods and to define critical design principles both for research itself and also for creating and using digital tools for different aspects of the research process.

The workshop will focus on digital cultural empowerment in particular looking at the ways in which humanities participants are empowered through using visualisation outside traditional domains, and harnessing the power of digital technologies to explore culture outside the confines of textual, linear narrative and traditional publication. We will focus on visualisation as the representation of knowledge in a multiplicity of new, engaging and challenging approaches across humanities disciplines.

Answers to research questions are being represented in innovative ways and we are interested in how these differing approaches are applied across the humanities without losing the specificity of their respective domains:

- How do historians, linguists, geographers, and art historians use visualisation?
- How are archaeologists integrating innovative visual methodologies in their research and publication?
- How does visualising literature augment and sustain research practices?
- How can deep mapping increase our understanding of historical processes?

Over the past two years, workgroups in NeDIMA have carried out an extensive survey of digital practitioners and their practices related to the use of visualisations and the representation of space and time. This work through a formal survey and through targeted workshops has been extensively mined and collated and the findings will be presented, shared, refined and form a guiding instrument for practise in these fields. This guide is the starting point of this workshop and its results will be shared and enriched by the participants.

Objectives:

- To bring together researchers with specific interests in method development considering the whole research process;

- To capture and share the experiences and progress of researchers developing or using visual methods either as an endresult or as part of the research process in different disciplines;
- To consider scholarly communication and to constantly redefine how we read visualisation and for what purpose.

Workshop organization:

Based on the survey and past workshops, we gathered evidence and have come to understand how different research questions can arise from these methods in terms of open data, collaboration, remediation, space, performance, impact and outreach and to that end, we welcome participation from researchers with innovative questions and methodologies to share in the workshop. This workshop will appeal to those who are beginning to experiment with visualisation as well as those with a deeper experience with visualisation, temporal, and spatial issues.

Participants are expected to prepare themselves by considering the following questions:

- Data capturing: How is data collected?
- Data modeling /data selections / data quality : How do we ensure consistency and high quality?
- Representation/visualisation /complexity vs "easy to read" : What do you choose to represent?
- Analysis : What does your reading of visualization bring to you?

- Purpose of visualization /Audiences: How do we make different representations for different audiences?

We propose to use 4 case studies as an inspiration/stimulation for sharing, discussing and summarizing participants' experiences:

- 1. Datamining the Nordic Folklore database faceted browsing aligned to research questions 2. Visualisation of time/space 3. Literature re presenting the text 4. Dataviz: Art or Representation of knowledge

This is a truly inter and transdisciplinary approach delivered by an international team that will highlight the variety of modes through which visualisation as a tool can facilitate exploration, knowledge creation, and understanding for researchers.

Target audience:

This workshop will target researchers with specific interests in methods empowering digital culture through using visualisation, particularly outside traditional domains. We will also be targeting invitations and mining the NeDIMA network's contacts from participants in NeDIMA workshops, specifically time/space and infoviz workgroups, the DARIAH network and attempting to engage new participants who are attending DH through more broad invitation on mailing lists and discussion forums. We will also use our contact with UCLA (Johanna Drucker at UCLA) as well as the CESTA lab at Stanford (Nicole Coleman, Zephyr Frank, Karl Grossner).

We will begin advertising the availability of the workshop upon workshop acceptance and will make it available until target is attained. If there is more demand than anticipated we will establish a waiting list and can possibly deliver an additional workshop in the morning.

We anticipate 30 participants based on past workshop attendance at DH2012 using first come first serve principle. Registration will be done via the DH2014 registration website.

Detailed program Tuesday 8th of July 2014

13.30 Session I (1 hours):

- Case Study/Provocation: Datamining the Nordic Folklore database
- Case Study/Provocation: Visualisation of time/space.
- Case Study/Provocation: Literature reflection on current trends in visualization

- Case Study/Provocation: From aesthetics to data visualization, is there a bridge between art and visualization?
- Instructions for the work in break out group
- One document per group of 6 answering the questions.
- Aiming at 5 groups.

14.30 -15.30 Breakout and discussions (1 hour)

16.30-17.30 Session II 1h

Reporting back from discussions Conclusions and further steps

Turn into an ongoing interactive document?

18.00-18.40 Joint walk to restaurant

Workshop leaders

Fredrik Palm

email: fredrik.palm@humlab.umu.se mobile: +46 70 3364438
Fredrik Palm is an expert in databases, GIS, visualization, webdevelopment and have been working in HUMlab since 2001. He has good management skills and been involved in writing several research proposals in the EU frame programmes for Research and Development (FP6 and FP7).

Fredrik initiated the proposal of the QVIZproject and was the assistant project manager of the FP 6 project (20062008). Fredrik has an Informatics degree and a teaching degree in History, Geography, Religion and Social sciences. Fredrik's role in HUMlab is to act a translator/broker between the needs of researchers and the potential of information technology.

Furthermore he has been involved as development coordinator in the BIOMAP (20082010), SHIPS (20092011) and the SEADproject (20082013) and "Digital publication of Rock-Carvings at Nämforsen" (20112013).

He is also involved in the DIABASproject (2003,20102013) and the ERCfounded project "Mapping the Jewish Sects of the Byzantine Empire" (20122013). He is also involved in the CITIZMAPproject founded by Vinnova (20112013). Fredrik Palm is also leader of a working group 2 "Visualisation" in Network for Digital Methods in the Arts and Humanities - NeDiMAH.(20112014)

Shawn Day

shawn.day@ucc.ie
+353 (0)83 0024264

Shawn Day blends the aesthetic and informative as an entrepreneur, digital historian, economist. He lectures at University College Cork, Queen's University Belfast and Trinity College Dublin, in Digital and Medical Humanities and Social Computing.

His personal research explores the social and economic circumstances of the nineteenth century retail liquor trade and it's impact on family. He applies digital, spatial and social network analysis to the relationships between credit, respectability, and order in the Victorian community. Recent articles have examined the social dimensions of the Victorian public mental hospital using GIS and statistical modeling tools. Shawn has been involved in a number of successful and innovative digital humanities projects. These include large manuscript census databases in the 1871/1891 census project (University of Guelph), the national TAPoR text analysis portal project, the Canadian Network for Economic History (CNEH) and the Network for Canadian History and the Environment (NiCHE).

Shawn has blended his background in management economics with an entrepreneurial ethos to found a number of successful software development ventures in Canada and applies this experience to academic activities.

Orla Murphy

o.murphy@ucc.i
+353 (0)21 4902591

Orla Murphy is a lecturer in the School of English at University College Cork, in the Irish national, inter-institutional Digital Arts and Humanities PhD program, and co coordinator of the MA in Digital Arts and Humanities at UCC. She is interested in digital pedagogy and in working to create an online MA in

Digital Cultures, launching in September 2014. Her research is focused on intermediality, on how the text is, was, and will be transmitted; how we read, represent, and share knowledge in new networked and virtual environments. She is co chair with Fredrik Palm, HUMlab Sweden, of the information visualisation working group in NeDiMAH.eu, (Network for Digital Methods in the Arts and Humanities) and vice chair of the EU CoST (Cooperation in Science and Technology) CoSCH.info working group on algorithms and representing 3D; she is also Irish representative on the CoST transdomain action on Gender in Science and Technology.

Nicolas Thély

nicolas.theley@univ-rennes2.fr
+33 681 86 87 81

Nicolas Thély is professor in Digital Humanities at Université de Rennes 2 (France). He teaches digital art and aesthetics. As an art critic and theorist he has published *Vu à la Webcam* (Essai sur la Web-intimité), (2002), *Corps, Art Vidéo et Numérique* (2005) and *Mes Favoris* (2007). Between 2007 and 2011, he has headed the Basse Définition research project design to show how algorithms such as PageRank, MP3, MPEG, and GIF have restructured the sensorial environment and creative power of contemporary artists. Since 2011, Nicolas Thély use digital tools and quantitative methods to debate in aesthetics about Art criticism. Last October, He organised with Alexandre Serres and Olivier Le Deuff THATCamp Saint-Malo (DH and design). He is professor referent for the digital humanities platform of the Maison des Sciences de l'Homme en Bretagne (MSH-B).

His last publications : *Search terms : Basse déf.* (dir.), éditions B42 ; *Le Tournant numérique de l'esthétique*, collection Art, pensée & Cie, Publie.net (2011) ; *"Rôle et enjeux du design graphique"*, THATCamp Paris 2012, *Non-actes des non-conférences des humanités numériques*, éditions de la Maison des Sciences de l'Homme, Paris, 2012 ; *"Archiver le Web"*, Read/Write Book n°2 (Introduction aux humanités numériques), dir. Pierre Mounier, Open Edition Press, Paris, 2012.

Prosopography Workshop

Quamen, Harvey

hquamen@ualberta.ca
University of Alberta

Crompton, Constance

constance.crompton@gmail.com
University of British Columbia, Okanagan

Hjartarson, Paul

paul.hjartarson@ualberta.ca
University of Alberta

This workshop is co-sponsored by the Prosopography Working Group, Canadian Writing Research Collaboratory Prosopography Group, Editing Modernism in Canada (Alberta collaborative) and the Humanities Computing Program at the University of Alberta.

Biographies

Harvey Quamen is an Associate Professor of English and Humanities Computing at the University of Alberta, where he contributes to several digital humanities research teams, including the Canadian Writing Research Collaboratory and the Implementing New Knowledge Environments project. With the Editing Modernism in Canada cluster at the University of Alberta, he has helped develop a literary walking app, WatsonWalk, and is currently researching how data visualizations can be used archival searches and prosopographies. He teaches a course on "Databases for Digital Humanists" annually at the Digital Humanities Summer Institute at the University of Victoria.

Constance Crompton is an assistant professor of Digital Humanities at the University of British Columbia, Okanagan

Campus, with research interests in data modelling and curation, queer history, and Victorian popular and visual culture. She is co-director of Lesbian and Gay Liberation in Canada, an infrastructure pilot project of the Canadian Writing Research Collaboratory at the University of Alberta and a member of the Implementing New Knowledge Environments (INKE) modelling and prototyping team. Her work has been published in the Victorian Review, Nineteenth-Century Gender Studies, and the UBC Law Review.

Paul Hjartarson, Professor of English at the University of Alberta, researches primarily in the areas of modernism, print culture, and the digital humanities. With Harvey Quamen, he oversees the EMiC UA research group and leads the Editing the Wilfred Watson Archive Project. He contributed to volumes 2 and 3 of the History of the Book in Canada. With Gregory Betts and Kristine Smitka, he is co-editing Counterblasting Canada, an essay collection that assesses the importance of Marshall McLuhan's media theories for Canadian writers and artists. With Shirley Neuman, he is editing the letters Sheila and Wilfred Watson wrote one another between 1956 and 1961.

Workshop Details

Duration: half day
Budget: none

Brief Workshop Outline

This half-day workshop is designed to promote interest in—and share resources for building—prosopographies. In the most general sense, prosopographies are simply databases or encoded files containing information about people. The groups sponsoring this workshop are all involved in the building of prosopographies of various sizes and so have been pooling resources and developing best practices. This workshop is an extension of that community's work.

Although the prosopographical form dates from the 16th century, the computer's ability to search and store vast quantities of data has lead to a recent expansion of prosopographical work in a digital context. Always somewhat interdisciplinary in nature, prosopographical studies are described by scholar K.S.B. Keats-Rohan as being "an independent science of social history embracing genealogy, onomastics and demography" (*Prosopography Portal*). King's College London scholars John Bradley and Harold Short agree, but extend their definition of prosopography in ways that are central to this workshop. In their estimation, a "new-style" prosopography consists of a collection of so-called factoids: "assertions made by the project team that a source 'S' at location 'L' states something ('F') about person 'P'.... A factoid is not a statement of fact about a person.... Instead, each one records an assertion by a source at a particular spot about a person. Factoids may contradict each other" (8).

Prosopographies, then, are not the same as authority list records because prosopographies can often capture contradictory or untrue information—precisely the type of information that would be ruthlessly edited out of an authority list. Prosopographies can be fruitful resources for humanities research, however, because as they collect historical claims about people, they tie those claims to sources. Prosopographies can track rumour, myth, debate, and propaganda in ways that reveal the personal and political struggles that constitute history. Prosopographies, then, can not only promote new kinds of research questions, but can also track contentious research debates themselves.

Prosopographies and authority lists often work in conjunction with one another, and so the distinction between fact and factoid need not pose a practical conundrum. Indeed, the workshop will look at a variety of "open data" authority lists, including the giant Virtual International Authority File (VIAF). Stored in a relational prosopographical database, for example, an authority list might constitute just one table amid other tables designed to store assertions and their respective links to sources.

Workshop Agenda

The 3-hour workshop will be divided into two sections: Part One will last approximately an hour and be devoted to discussion about, and brainstorming of, the types of data that practicing and aspiring prosopographers need to capture. To frame the discussion, we'll make available online some of the theoretical readings (including works by Keats-Rohan, Bradley and Short, Alison Booth, etc.), and participants will be invited (but not necessarily required) to read them before the workshop begins. Workshop participants will come away from this part of the day with a workshop-generated bibliography of prosopography-related resources and scholarship. (Note: this discussion could be enhanced by participants who might have participated in a workshop proposed by John Bradley on the Ontology of Historical Persons).

Part Two (approximately two hours in length) will be devoted to the topics of data capture and long-term storage. We will examine the use of low-tech data entry tools (like Excel spreadsheets and XML documents) and then we will consider long-term database storage for these materials. With sample datasets, we'll demonstrate data design issues in both a relational database (using MySQL) as well as a graph database (using Neo4j). Both MySQL and Neo4j are free, open source tools.

We'll introduce workshop participants to the two distinct, but related, query languages used by these databases: Structured Query Language (SQL) and the graph-database variant, Cypher. A three-hour workshop will not make query experts out of the workshop participants, but they will be introduced to these languages and will be better able to judge which platform might hold the most promise for their respective projects.

Finally, we will gauge interest in ongoing prosopography work among participants and in developing an international Prosopography Working Group. The workshop organizers currently represent a small Canadian Prosopography Working Group; however, with an online presence, our working group will be able to facilitate the sharing of ideas and resources across national boundaries more effectively. Should there be sufficient interest, our ultimate goal is to hold a conference to help formalize this research and share it among interested groups.

Participant Preparation

Advance Readings (optional)
Loading MySQL and Neo4j (optional)

Works Cited

- Bradley, John, and Harold Short.** "*Texts into Databases: The Evolving Field of New-style Prosopography*." *Literary and Linguistic Computing* 20 (2005). 3-24.
MySQL. <<http://www.mysql.com>>. Neo4j. <<http://www.neo4j.org>>. *Prosopography Portal*. <<http://prosopography.modhist.ox.ac.uk/>>. Virtual International Authority File. <<http://viaf.org>>.

Leveraging Web Archiving Tools for Digital Humanities Research and Digital Exhibition

Reed, Scott Brian

scott@archive.org
Internet Archive, United States of America

Summary:

Web archiving is an important part of the digital preservation field. While most are familiar with the Wayback Machine available at archive.org, less are aware that there are a number of tools and services developed for organizations and individuals to create their own web archives, including the capability to search and analyze large data sets built around the WARC file format, an ISO standard for web archiving. In addition, web archives provide permanent URLs for citation and can show how a website has changed over time at a single URL, even if no longer available on the live web. In short, web archives can save a researcher's life and provide very necessary preservation tools for archivists to manage content that is only posted on the web.

This workshop will introduce participants (15-20) to basic web archiving concepts and challenges. Using the Archive-It (www.archive-it.org) web application, participants will have a hands-on opportunity to build a collection of content archived from the web, which can include their own organization's web presence, social media, digital exhibitions, data sets, or topical content publicly available on the web. Following the workshop participants will have a searchable archive available to them, including the option of downloading WARC files for long term preservation or research.

The target audience for this workshop includes interested humanities scholars researching the web and professionals responsible for digital library service or digital archives. No prerequisite knowledge of or experience with web archives is necessary, and the session does not require any programming or advanced technical knowledge of the web. The workshop will not be oriented towards those with deep knowledge of web archives or the WARC format, although there could be time allotted to a demonstration of another web archiving tool or project related to digital humanities and web archiving and this should be specified in the CFP (see below).

CFP:

In order to make the most of the 3 hour workshop and ensure that the curriculum is tailored toward participant interest, a CFP will be requested for all interested persons. It should include:

- Description of participant research interest or professional projects.
- Description of prior experience with using web archives or their own web archiving (if applicable).
- 5 to 10 websites to be archived as part of 1 or more collections of content, and links to the Robots.txt files if applicable. More information is here: <https://webarchive.jira.com/wiki/display/ARIH/Robots+Exclusion+Protocol#RobotsExclusionProtocol-Whatistherobotsexclusionprotocol?>

With permission from participants, URLs will be crawled as a test (no data archived) prior to the workshop so post crawl reports can be analyzed as part of the workshop curriculum.

The CFP process is not intended for competitive review but to ensure relevancy and preparedness of participants. It should be received at least 2 weeks before the workshop. CFPs will be reviewed by the instructor and Kristine Hannah, Director of Web Archiving Services at the Internet Archive.

Cost and Equipment Required:

There will be no additional costs associated with the workshop. It will require a meeting room with wireless internet and a projector with screen. Participants will bring their own wi-fi enabled laptop computers and there should be sufficient power outlets.

Agenda Outline:

15 minutes

Welcome and participant introductions including overview of content being archived

25 minutes

Overview of web archiving, including:

- history of web archiving
 - common software tools (Heritrix, Wayback)
 - overview of WARC file format
 - overview of Internet Archive and Archive-It
- 60 minutes**
Hands-on Archive-It web application training
- creating a collection
 - adding and scheduling seeds to be crawled
 - starting and monitoring (crawls)
 - modifying the scope of the crawl
 - understanding Robots.txt
 - analyzing post crawl reports
 - quality assurance, including addressing web archiving limitations and challenges

30 minutes

Other Web Archiving tools (including WAIL)

30 minutes

Understanding the web archiving life cycle, including:

- open source tools for researching WARC files
- future steps for web archiving
- sharing and reporting on future research projects utilizing web archives (group activity/share)

20 minutes

Flex time, conversation, and breaks

About the instructor:

Scott Reed has worked as a Partner Specialist with the Internet Archive since 2012, primarily supporting organizations and researchers using Archive-It to build collections of web content. In addition he is a volunteer with the GLBT Historical Society in San Francisco, CA . Prior to his work with Internet Archive, he has worked in various positions as a digital literacy and media instructor and project assistant for non-profit organizations and academic departments in California including the Feminist Studies department of the University of California, Santa Cruz.

Multilinguality in historical documents – challenges and solutions for digital humanities

Romary, Laurent

INRIA & HUB-IDSL & Dariah, Germany

Dipper, Stefanie

RUB

Bubenhofer, Noah

TU Dresden

Vertan, Cristina

Uni Hamburg

Recently, the collaboration between the Language Technology community and the specialists in various areas of the Humanities has become more efficient and fruitful due to the common aim of exploring and preserving cultural heritage data. It is worth mentioning the efforts made during the digitisation campaigns in the last years and within a series of initiatives in the Digital Humanities, especially in making old manuscripts and prints available in the form of Digital Libraries.

The availability of old texts on-line produced a revolutionary shift in the way how such objects are analysed. They are no longer restricted to a small number of specialists, knowing the language of the document but to broader groups with various requirements:

- non-expert users who would like to know what the document is about, understand the main topics, localise places, persons. These users have no or very little knowledge of old languages, and usually are less familiarised with toponyms

- (especially when these belong to geographical spaces unknown to the user);
- researchers of neighbour fields, who often have only minimal knowledge of the language but considerable knowledge of the historical context and might be familiarised with historical toponyms and proper names;
 - students and researchers specialising in historical data, who have the required language skills but still can profit from additional information accompanying the texts.

These considerations imply that the storage and visualisation of old texts should be accompanied by a collection of tools empowering the text with suitable information and making it understandable for different user groups. Such tools usually involve automatic language processing methods. In contrast to processing of modern texts, for which language technology made a huge progress in the last years, automatic processing of old texts is still problematic mainly because:

- Historical language data is sparse. First, compared to the wealth of documents written in modern languages, there are only few documents available for historical languages. Second, transcribing old manuscripts often requires expert knowledge. Third, due to the absence of a standard language, historical language variants differ in spelling, morphology, syntax, and lexical semantics from each other.
- Texts are often multilingual, consisting of mixtures of different languages, such as single words or phrases or entire sentences written in Latin that are intermixed with passages written in the actual language of the text. In case of texts from areas with rich cultural mixtures (e.g. Balkans), one can find in addition paragraphs in "exotic" local languages.

The focus of this workshop is on the second aspect. We think that the challenges posed by multilinguality should be tackled by adapting existing multilingual language resources and tools, and, where necessary, by providing training data in the form of corpora or lexicons for a certain period of time in history.

The aim of this workshop is to bring together researchers working in this interdisciplinary domain as well as specialists in machine translation and multilinguality working with languages with sparse resources, to analyse problems and brainstorm solutions in order to implement machine (-aided) translation and processing for (multilingual) historical texts. We envisage also networking with European activities in Digital Humanities like CENDARI, CLARIN, DARIAH.

Topics of interest include but are not limited to:

- character-level Machine Translation (MT) for normalisation
- historical and modern data as comparable corpora (methods for extraction parallel segments from translations or new editions in modern language)
- historical texts in different languages as parallel or comparable corpora
- MT for translation between language versions
- OCR for multilingual documents
- word- and/or paragraph-level language identification
- crosslingual retrieval in historical documents
- ontologies as language-independent interfaces between collections of historical texts
- particularities of multilingual historical texts and challenges for IT

Cristina Vertan, University of Hamburg
 Research Group „Computerphilologie“, University of Hamburg, Vogt-Kölln Strasse 30, 22527 Hamburg, Germany
 Cristina.vertan@uni-hamburg.de, Office: +49 40 42838 2319
 nats-www.informatik.uni-hamburg.de/CristinaVertan
 Cristina Vertan is senior researcher at the University of Hamburg. Her principal research fields are Machine Translation, Digital humanities, Crosslingual retrieval and less-resourced languages. She organised several workshops at important conferences (LREC, RANLP) about using language technology for cultural heritage and historical languages. Se his founding member of th SIGHUM special ACL-interest group in „Digital Humanities“ and co-organiser of the LATECH-2014 (Language Technology for cultural Heritage Social Sciences and Humanities) collocated with EACL 2014. Recent research activities iclude extraction of parallel corpora from historical

translation, one paper being accepted at the Digital Humanities Conference 2014.

Full contact information for all workshop leaders, including a one-paragraph statement of their research interests and areas of expertise;

Stefanie Dipper, Ruhr-Universität Bochum
 Sprachwissenschaftliches Institut, Ruhr-Universität Bochum, D-44780 Bochum,
 dipper@linguistics.rub.de, Office: +49 234 32-25112
 Stefanie Dipper is Professor of Computational Linguistics at Ruhr-University Bochum, Germany. She has worked on annotation formats, corpus tools, and corpus-based methods for many years. Her primary interests are in automatic analysis of historical texts, including normalization of historical spelling, POS and morphological tagging, and in methods for comparing and clustering historical dialects. She is PI of a DFG-funded project that deals with creating and analyzing a corpus of historical dialects, and Co-PI of two DFG-funded projects for creating reference corpora of historical German.

Noah Bubenhofer, TU Dresden / UZH Zurich
 TU Dresden, Institut für Germanistik, Professur für Angewandte Linguistik, Mommsenstr. 13, D-01062 Dresden, noah.bubenhofer@tu-dresden.de, Office: +49 351 46 33 82 19, Mobile D: +49 170 901 17 94,

Mobile CH: +41 76 330 66 15
 Web: www.bubenhofer.com , linguistik.zih.tu-dresden.de

Dr. NOAH BUBENHOFER is a member of the academic staff at the Chair of Applied Linguistics, Technische Universität Dresden and head of the recently opened Dresden Center for Digital Linguistics. In addition, he is co-founder of SEMTRACKS, the „Laboratory for Computer Based Meaning Research“. Since 2014, Noah Bubenhofer is a guest researcher at the Institute of Computational Linguistics at the University of Zurich. In his PhD-thesis „Muster an der sprachlichen Oberfläche“ (patterns at the linguistic surface), he developed corpus linguistic methods for discourse and cultural analysis. As a linguist, he is mainly interested in computer based semantic text analysis and the relation between text and discourse, society and culture. In the project „Tracking Meaning on the Surface“ categories were modelled for the description of semantic imprints using a data-driven approach. In doing so, the project explored possible applications of these models for the semantization of the Internet and the methodology of social sciences and cultural studies. Noah Bubenhofer is also co-leader of the project “Text+Berg digital” (www.textberg.ch) where a series of yearbooks by the Swiss Alpine Club (SAC) is being digitised and transformed into a deeply annotated corpus.

Laurent Romary, laurent.romary@inria.fr

Inria & HUB
 Institut für deutsche Sprache und Linguistik
 Philosophische Fakultät II
 Humboldt-Universität zu Berlin Unter den Linden 6
 D-10099 Berlin

Laurent Romary is Directeur de Recherche INRIA, France and guest scientist at Humboldt, University in Berlin Germany. He carries out research on the modelling of semi-structured documents, with a specific emphasis on texts and linguistic resources. He received a PhD degree in computational linguistics in 1989 and his Habilitation in 1999. During several years he launched and directed the Langue et Dialogue team at Loria in Nancy, France and participated in several national and international projects related to the representation and dissemination of language resources and on man-machine interaction. In particular coordinated the MLIS/DHYDRO, IST/MIAMM and eContent/Lirics projects. He has been the editor of ISO standard 16642 (TMF – Terminological Markup Framework) and is the chairman of ISO committee TC 37/SC 4 on Language Resource Management. He has been member (2001-2007) then chair (2008- 2011) of the TEI (Text Encoding Initiative) council. In the recent years, he lead the Scientific Information directorate at CNRS (2005-2006) and established the Max-Planck Digital Library (sept. 2006-dec. 2008). He currently contributes to the establishment and coordination of the European Dariah infrastructure.

DARIAH-EU VCC2 Workshop on Innovative Teaching Methods and Practices in Digital Humanities

Scholger, Walter

walter.scholger@uni-graz.at
University of Graz, Austria

Clivaz, Claire

claire.clivaz@unil.ch
University of Lausanne, Switzerland

Tasovac, Toma

ttasovac@humanistika.org
Belgrade Centre for Digital Humanities, Serbia

Outline

The growing interest in Digital Humanities has resulted in an increasing number of individual courses, modules and even degrees covering a broad range of topics at the cross-section of humanities and ICT-based methods. Despite numerous efforts to formally train students and researchers in the wide-ranging field of Digital Humanities, "scholarship in this area has tended to focus on research methods, theories and results rather than critical pedagogy and the actual practice of teaching" (Hirsch 2012). The increasing recognition and institutionalization of digital humanities in academic departments seems to have been coupled with the traditional and highly problematic division of labor between research and teaching as two antagonistic activities, of which only the former significantly contributes to the advancement of academic careers, including tenure. For that reason one of the most important questions facing our field today remains whether we can — in both theoretical and practical terms — pursue not only new ways of thinking about the humanities, but also new ways of teaching and interacting with students as part of our core professional activity.

With this workshop, we would like to motivate colleagues with an interest or actual experience in using innovative methodological approaches to teaching Digital Humanities to showcase precedent-setting developments, and encourage the participants - and, consequently, the DH community at large - to share their thoughts and ideas on how the development of a digital pedagogy for digital humanities should proceed. We would like to explore how DH processes and challenges that emerge out of building tools, developing projects and using computational methods to analyze data, influence ways in which to engage students and whether this engagement can, should and does in fact lead to epistemological turns and pedagogical transformations.

The proposed workshop is part of the efforts by the Virtual Competency Centre 2: Research and Education Liaison of the European ESFRI initiative Digital Research Infrastructure for the Arts and Humanities (DARIAH) to promote and support the use of research data and ICT methods and technologies in the Humanities¹. The workshop will therefore also highlight the importance of pedagogical considerations in the process of building infrastructures: Unless a great deal of careful thought is given to education and outreach, as well as the integration of digital infrastructures into DH curricula and training activities, digital infrastructures will always run the risk of fossilizing before even gaining momentum.

The co-leaders of DARIAH-EU VCC 2's task group Training and Education - Claire Clivaz, Walter Scholger and Toma Tasovac - will serve as workshop leaders. The program committee for the selection of presentations is drawn from experts within DARIAH-EU.

The workshop will be divided into two 3 hour sessions:

Session 1 (Practical): Showcases and best practices for teaching DH

In the morning session, participants will have the opportunity to present their ideas and/or actual teaching methods and materials. The contributions presented will be collected through

a workshop-specific call and selected by a program committee. Our aim is to attract not only long-time practitioners of DH but also recent adopters with innovative ideas and methods to present a showcase which will motivate other participants to reuse and promote such contributions in their own field, or spark the development of yet more original methods.

Session 2 (Colloquium): Challenges in DH pedagogy

The afternoon session will provide a forum for the participants to discuss the most prominent challenges and issues in (digital) teaching of DH. Drawing on the impressions and ideas presented during the Practical Session, the participants will assess what the main challenges in DH are and agree on specific issues and topics for further discussions which will be taken to breakout sessions moderated by the workshop leaders.

In addition to discussing solutions to common challenges, participants will also enter in a critical discussion of the necessary next steps in promoting DH through digital pedagogy. As a starting point, the workshop leaders will present objectives that were formulated in DARIAH and discuss their value and implementation with the participants. These objectives will be amended and modified by the contributions of the participants to reflect the interests and issues agreed upon by the community (i.e. the workshop participants).

Dissemination

A tangible outcome of the workshop will be the preparation of a report on Digital Teaching Methods and Practices in DH which will document the best-practice examples presented during the workshop, but also the issues - and, hopefully, solutions - raised during the discussions and participant-driven breakout sessions. In addition, the contributions to the workshop will be highlighted on a website that will also serve as a communication channel for the participants after the workshop.

The workshop would also serve as a kick-off for other efforts by DARIAH-EU to promote this topic on a larger scale: A series of national DARIAH workshops organised for promoting DH and especially innovative teaching methods for DH content, tailored to the specific situation in the hosting DARIAH member countries, will be held over the period of two years.

Workshop Leaders

Claire Clivaz [email: claire.clivaz@unil.ch / Tel. +41 692 2714]

is an Assistant-Professor in New Testament and Early Christianity and member of the board of the Laboratory of Digital Humanities and Cultures (Ladhu) of the University of Lausanne. She is in charge for the French part of Switzerland of a pilot project for a future Swiss DH center, under the lead of the DHLab Basel and with Bern University, and leads a DH seminar for PhD students and post-doc researchers, in collaboration with the DHLab EPFL. She is a member of the EADH committee and the DHSI board.

Walter Scholger [eMail: walter.scholger@uni-graz.at / Tel: +43 316 380 2292]

is the deputy head of the Centre for Information Modelling - Austrian Centre for Digital Humanities at the University of Graz (Austria). His main research areas are (digital) IPR and copyright issues, and the development of curricula and training modules for Digital Humanities contents.

Toma Tasovac [email: ttasovac@humanistika.org / Tel. +381 66 9373250]

is the director of the Belgrade Centre for Digital Humanities. He works on complex architectures in electronic lexicography, digital editions, and integration of digital libraries and language resources. He is equally active in the field of new media education, regularly teaching seminars and workshops in Germany, Eastern Europe, the Caucasus and Central Asia.

Organisational Matters

Scope

The workshop is conceived as a full-day workshop, distributed into two separate sessions as outlined above. If at all possible, we would ask to hold this workshop on July 7, since a meeting of the French-speaking DH community has been confirmed for July 8.

Participants

The target audience are peers with an interest in digital pedagogy, primarily Early Stage Researchers/Teachers who are recent adopters of Digital Humanities methods and/or in the process of designing new lectures with DH content.

We expect up to 40 participants and would limit the number of attendees to that amount, since the second half of the workshop will feature discussions and audience-driven breakout sessions.

Technical requirements

We expect participants and presenters to bring their own devices, but we will require a strong wi-fi connection to handle simultaneous access to the internet by the workshop's participants and sufficient access to power supply. For the presentations, we will require a data projector and a large canvas.

Workshop-specific Call

A call for contributions showcasing innovative teaching methods and/or concepts will be distributed through relevant mailing lists and the DARIAH communication infrastructure to reach a broad audience. The call would be issued immediately after acceptance of the contribution (March 18) and would be open for a period of roughly 3 weeks, up to April 4. Notifications regarding the acceptance of proposals and details on the presentation framework will be sent out by April 28.

We will ask for abstracts of up to 500 words, detailing contributions of 5-15 minutes. The length allocated to each contribution will be decided by the program committee, depending on the number of contributions and the strength of the proposal.

Upon acceptance, contributors will have the opportunity to distribute material (presentation slides, videos, ...) to the participants via the dedicated workshop website in order to give them a chance to contextualize their contribution beyond the scope of their allocated presentation time.

Program Committee

- Agiati Benardou (Greece)
- Marianne Huang (Denmark)
- Anne Joly (France)
- Matt Munson (Germany)
- Kristoffer Nielbo (Denmark)
- Johanna Puhl (Germany)
- Stefania Scagliola (Netherlands)
- Susan Schreibman (Ireland)
- Manfred Thaller (Germany)

Introduction to Starting and Sustaining DH Centers

Siemens, Lynne

siemensl@uvic.ca
University of Victoria

1. Introduction to Starting and Sustaining DH Centers

Description

Digital Humanities is growing in scale from a series projects undertaken by a couple of individuals to institutionally-based organizations, with larger budgets and mandates, beyond other factors. How can a group of interested researchers, academic

staff, librarians and other invested stakeholders work together to create such a centre? This half-day workshop will address this question by exploring issues related to scaling operations from the individual to a centre, determining the appropriate organizational model, developing the plan which situates the DH centre in the academic institution's and other stakeholders' mandates, communicating to administration to gain support and resources, structuring memorandums of understanding between partners, and other issues. It builds on centerNet's initiatives to support center startups with information and tools, which include the DHCenterStartUp listserv and resources page (<http://digitalhumanities.org/centerNet/resources-for-starting-and-sustaining-dh-centers/>).

Workshop leader

Dr. Lynne Siemens, Assistant Professor, School of Public Administration, University of Victoria and centerNet's Coordinator for Center Startups with centerNet.

Dr. Siemens is an Assistant Professor at the University of Victoria. Her interests include academic entrepreneurship, collaboration and teamwork with a focus on understanding methods and processes to facilitate collaborative research across distances, disciplines and organizational boundaries. Beyond publishing in these areas, she has taught workshops in Project Management at University of Victoria's Digital Humanities Summer Institute and University of Leipzig's European Summer School for Culture and Technology and serves as a management advisor for Implementing New Knowledge Environments (INKE), a Major Collaborative Research Initiative project. Dr. Siemens is also centerNet's Coordinator for Center Startups and coordinates the DHCenterStartUp listserv and the resources for starting DH centers webpage.

CenterNet is an international network of digital humanities centers. Part of its mandate is to support the development of centers at a variety of institutions by sharing information, resources and expertise.

Contact Information

Lynne Siemens
School of Public Administration
University of Victoria
Victoria, British Columbia Canada
V8W 2Y2
(250) 721-8069
siemensl@uvic.ca

Target audience and expected number of participants

The target audience is individuals interested in starting a Digital Humanities center at their institution. Expected number of participants is 15-20. (These numbers are in line with previous offerings of similar workshops/talks.)

Intended length and format of workshop

The workshop will be a half-day with a combination of lecture and discussion. No special technical requirements or equipment is needed.

Kickstarting the GO::DH Minimal Computing Working Group

Simpson, John Edward

john.simpson@ualberta.ca
University of Alberta, Canada

Sayers, Jentery

University of Victoria, Canada

O'Donnell, Daniel Paul

daniel.odonnell@uleth.ca

University of Lethbridge, Canada

Gil, Alex

colibri.alex@gmail.com

Columbia University, USA

What is the purpose of this workshop?

Establish a mandate for the Global Outlook::Digital Humanities (GO::DH) Minimal Computing Working Group so that it can serve the DH minimal computing community as it wants and needs to be served.

What questions will the workshop answer?

What is the current state of minimal computing in the DH community?

What will be done to further support minimal computing users within the DH community?

Why this workshop? Why should the DH community care about minimal computing?

With machines like the Tihane2, a 33.86 petaflop computer featuring 3.12 million cores and only the most recent machine to best the highperformance computing (HPC) Top 500, the current push within large parts of the DH community to get access to and ultimately use such machines (cf. Bonnett 2009; Leetaru 2012; Terras 2009; The NEH High Performance Computing Collaboratory), and the desire to do big things with a whole lot of data and slightly less powerful machines (cf. The Digging into Data program; HuNI; ARC; CWRC; various OCLC initiatives; <insert acronym of your choice here>) why the DH community should pay any attention to minimal computing certainly needs to be addressed.

The GO::DH Minimal Computer Working Group uses "minimal computing" to capture both the maintenance, refurbishing, and use of machines to do DH work out of necessity and the use of new streamlined computing hardware like the Raspberry Pi or the Arduino microcontroller to do DH work by choice. This dichotomy focuses the group on computing that is decidedly not highperformance and importantly not firstworld desktop computing. By operating at this intersection between choice and necessity minimal computing forces important concepts and practices within the DH community to the fore. In this way minimal computing is also an intellectual concept, akin to environmentalism, asking for balance between gains and costs in related areas that include social justice issues and demanufacturing and reuse, not to mention rethinking highincome assumptions about "ewaste" and what people do with it.

Interest in minimal computing can already be seen via workshops at places like Princeton, Carleton, McGill, MITH, Victoria, and HASTAC, each of which is just a small drop in the bucket compared to the growing Internet resources available for such project. But this is just the side of minimal computing that currently has enough cache and geekchic to have caught the momentary attention of a sliver of the Internet. As became apparent to the participants of the INKE conference held in Cuba in 2012, for roughly 60% of the world minimal computing is a fact of life rather than a tweeteworthy hobby and very little is known about this Still, DH practitioners facing these conditions are finding ways to overcome these barriers in ways that are at once smart and sensible.

Bringing together those who do minimal computing by necessity with those who do it by choice stands to benefit not only those with a stake in minimal computing by facilitating knowledge and expertise exchange but the DH community as a whole by shining a spotlight on a large portion of humanities

work that is currently going unnoticed, enabling further research to take place in these areas.

Why a DH2014 workshop?

The Minimal Computing Working Group will operate as an extension of GO::DH, which is in turn a special interest group of the Alliance of Digital Humanities Organizations. Given this pedigree holding the workshop that will craft the overall direction of the working group at the annual conference of ADHO makes good sense.

Who is the target audience and how many attendees are expected?

The workshop targets two audiences: those who do DH related minimal computing by choice and those who do it of necessity. Those who do it by choice are by far the smaller portion of the global DH community but are also the most likely to be able to attend the event and we would be delighted if we were able to achieve a 50/50 split amongst the attendees across these two groups. Given that this workshop has a broad general appeal and offers a greater opportunity for participation both at the workshop and afterwards we're hopeful that we will be able to draw 40 participants and allow about half of these to present a lightning talk.

What will be the format of the workshop?

The workshop will be divided into three distinct components arranged in an order that will allow all attendees to build their background knowledge and contribute:

1. A series of lightning talks (25 minutes) about current research or work being done with or in a minimal computing environment. These would be drawn in advance with a CFP. Those unable to attend the workshop but wishing to present would be invited to share videos.
2. A focused brainstorming session directed at collecting ideas and projects that the Minimal Computing Working Group or its members should consider pursuing. It is hoped that some form of participation will be open to those not on site, but this will depend on the infrastructure that is
3. The selection of a set of tasks, directives, and/or projects that the minimal computing working group will coordinate and support. These will follow directly from the previous stages but these might look something like programs to:
 - provide training to the DH community to use minimal computing tools
 - share/ship computing resources to areas that might better use them
 - track hardware and software use in the humanities on a global scale
 - provide or recommend packages of hardware and software that are effective and proven

How long will the workshop be?

We are asking for a halfday workshop on the assumption that this will allow three hours plus breaks to complete the outlined program by roughly dividing it into one hour sections.

What will be the cost of the workshop?

There should be no cost to the workshop or the conference as a whole beyond the provision of audiovisual resources requested below. If it turns out that there is a cost for those that this workshop would need to bear then we will look for funding or go without.

What timeline will the workshop be organized around?

Notice of acceptance.	Announce workshop on GO::DH listserve, Humanist, Twitter, and via the various member organizations of ADHO.
May 1	Camera ready, single page, submissions that outline current work or research with minimal computing. Since we are looking to provide a synopsis of current activity in the field and around the world our preference will be to accept or accept with revisions as many submissions as possible.
May 15	Acceptance notifications.
June 16	Submission of slides/videos and final one page summaries.
June 23	Distribution of PDF bound summaries.
July 25	Distribution of lightning talk videos on GO::DH site.

Who are the workshop organizers?

John Simpson will be the principal organizer and facilitator of the workshop. His work will be supported by Jentery Sayers, who is the other Minimal Computing Working Group cochair, and Dan O'Donnell, current chair of GO::DH, and Alex Gil, GO::DH vicechair.

John Simpson

john.simpson@ualberta.ca

Cochair of the GO::DH Minimal Computing Working Group, John is a Postdoctoral Fellow at the University of Alberta. He works with INKE and Text Mining & Visualization for Literary History doing research into the intersection of the Humanities with the Semantic Web. His research interests include Philosophy of Science & Technology, Game Theory, and the exposure and expression of difference in digital media. He is an instructor at DHSI 2014 on programming for humanists who have never done it before.

Jentery Sayers

jentery@uvic.ca

CoChair of the GO::DH Minimal Computing Working Group, Jentery is Assistant Professor of English at the University of Victoria. His research interests include comparative media studies, digital humanities, AngloAmerican modernism, computers and composition, and teaching with technologies. He is a member of INKE and is the founding director of the UVic Maker Lab. He is a past instructor at DHSI and will again be part of Physical Computing and Desktop Fabrication for Humanists in 2014.

Dan O'Donnell

daniel.odonnell@uleth.ca

Chair of GO::DH, Dan is Professor of English at the University of Lethbridge. His research interests include Old English language and literature, the history of the book, editorial and textual scholarship, humanities computing, and reception-oriented criticism.

Alex Gil

ag3339@columbia.edu

ViceChair of GO::DH, Alex is Digital Scholarship Coordinator, Humanities and History Division, Columbia University Libraries. His research interests include twentiethcentury Caribbean literature, critical theory, digital humanities, textual studies, book history, new media theory.

References

Bonnett, John (2009). "HighPerformance Computing: An Agenda for the Social Sciences and the Humanities in Canada." *Digital Studies / Le Champ Numérique* 1, no. 2 (June 16, 2009). www.digitalstudies.org/ojs/index.php/digital_studies/article/view/168.

"Diggng Into Data > Home." Accessed February 19, 2014. <http://www.diggingintodata.org/>.

"High Performance Computing Collaboratory | National Endowment for the Humanities." Accessed February 19, 2014. www.neh.gov/divisions/odh/institutes/highperformance-computingcollaboratory.

"Home | TOP500 Supercomputer Sites." Accessed February 19, 2014. www.top500.org.

Leetaru, K.H. (2012) "Towards HPC for the Digital Humanities, Arts, and Social Sciences: Needs and Challenges of Adapting Academic HPC for Big Data." In 2012 IEEE 8th International Conference on EScience (eScience), 1–6, 2012. doi:10.1109/eScience.2012.6404439.

"Personal Computers (per Capita) Statistics Countries Compared NationMaster." Accessed February 14, 2014. www.nationmaster.com/graph/med_per_com_percapmedia-personalcomputerspercapita.

Terras, Melissa M. (2009) "The Potential and Problems in Using High Performance Computing in the Arts and Humanities: The Researching EScience Analysis of Census Holdings (ReACH) Project." 3, no. 4. www.digitalhumanities.org/dhq/vol/3/4/000070/000070.html.

"Statistics." Accessed February 19, 2014. www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx.

www.diggingintodata.org

huni.net.au

idhmc.tamu.edu/projects/arc

www.cwrc.ca

oclc.org

www.conceptlab.com/deadmedia

digitalhumanities.princeton.edu/events

blogs.carleton.edu/dh/projectgallery/tweetingthedigital-humanitieswitharduinoraspberrypi

digihum.mcgill.ca/blog/2012/01/15/billturkelarduinofor-humanistsworkshop

mith.umd.edu/engl668k/?tag=raspberrypi

maker.uvic.ca

www.rpiforlibs.info

Instructables, a popular DIY website made up almost entirely of contributions from its user community, features over 2370 arduino projects and over 270 Raspberry Pi projects.

We make the claim that the majority of humanists have only limited access to computing power by looking at the website NationMaster ("Personal Computers..." 2014) and noting that roughly 60% of countries have less than one computer for every ten people. It should be noted that while this information is drawn from 2005 the most recent report from the International Telecommunications Union ("Statistics" 2014) which covers from 2005 to 2013 presents a picture of slow growth since then resulting in a picture that isn't all that much brighter.

My Very Own Voyant: From Web to Desktop Application

Sinclair, Stéfan

McGill University, Canada

Rockwell, Geoffrey

University of Alberta, Canada

Context

Voyant Tools (voyant-tools.org) is a web-based reading and analysis environment for digital texts. Users can create their own corpus of texts to work with by pointing to URLs or uploading files in a variety of formats (plain text, XML, HTML, PDF, MS Word, RTF, etc.). Voyant allows users to

navigate between macro views of the corpus (e.g. a word cloud visualization of the entire corpus) and micro views (e.g. a reading individual occurrences of a specific term in context). The default interface provides access to eight tools for reading texts and studying frequency and distribution data, and many more tools are available in various pre-defined or user-defined "skins" (a layout of tools that interact).

Using the hosted version of Voyant has several advantages. For instance, Voyant can be accessed immediately through most modern browsers and no software installation or configuration is required (some of the tools need plugins for Flash or Java, but not the ones in the default skin). Similarly, users are always accessing the most recent stable version of the web application, no actions are required by the user to update the software (typically, new builds with bug fixes and new functionality are released several times a month). A third benefit relates to Voyant's design as a networked tool where users can generate persistent URLs for their corpus and tools that can be stored, shared and embedded in remote sites.

The accessibility and convenience of the hosted version of Voyant also has disadvantages, however. Users often tell us they are wary of uploading texts to a server out of concern for possible copyright infringement or privacy issues. Consistent with its overall design priority of keeping things simple, Voyant does not currently have any user or access management system, so technically any uploaded content could be accessed by anyone (in practice, each uploaded corpus has a generated ID that it would be extremely unlikely to guess – so corpora are as hidden as users choose them to be). Still, with some material (protected responses to a survey that has required ethics approval, for instance), uploading content to a public, non-secure (non-SSL) site is not an acceptable option. A second potential source of frustration with the hosted version of Voyant is performance, especially with larger-scale text collections. The hosted server is configured to timeout after about 30 seconds, which can be insufficient when uploading, pre-processing and indexing a large text collection (the issue in some cases is not so much the scalability of Voyant as an application, but the constraints of a web application configured to support multiple concurrent users). Performance can also be an issue when running a workshop or tutorial with Voyant: 30 users hitting the same button at once can overwhelm the server (the current hosted deployment of Voyant runs in a ComputeCanada High Performance Computing facility but does not run as a cloud service that could smoothly expand capacity as on-demand). A final disadvantage with the hosted version of Voyant is that it requires reliable internet connectivity, which may be a mere nuisance at 30,000 feet, but a true problem at, say, a conference workshop with slow or unstable wireless. Performance, privacy, connectivity: three reasons among others to explore alternatives to using the hosted version of Voyant.

Workshop Outline

1) What is Voyant? (1 hour)

We will begin with a general introduction to Voyant within the broader landscape of digital text analysis. We will provide context and resources for working with digital texts. We will provide a brief overview of Voyant's user interface and discuss its strengths and weaknesses. We will suggest some other tools and techniques that may be of interest to workshop participants. This component of the workshop is like a mini-workshop on Voyant that will ensure a common base of familiarity with text analysis in general and Voyant in particular.

2) My Voyant (1 hour)

We will summarize some of the pros and cons of working with the hosted version of Voyant compared to running a local, stand-alone version. We will describe how to acquire a version of Voyant that can be run locally and then guide participants through running Voyant as a simple click-and-run desktop-style application. We will point out some of the most important technical aspects to be aware of, including how to understand the versioning system, where to find locally stored corpora, and how to respond to common errors.

3) Tweaking My Voyant (1 hour)

The nature of a web-application will be deliberately underemphasized in the previous component in favour of presenting My Voyant as a simple desktop-style application. This third component of the workshop will revisit the nature of My Voyant and explore ways that the web application can be tweaked to improve performance (memory settings in Java, server timeout settings, etc.) or otherwise fine-tuned (such as changing the default server port). The content will still be oriented around what's most relevant for a typical user of Voyant (not, say, a unix system administration guru). Finally, we will describe some strategies for deploying multiple instances of My Voyant for teaching/training purposes.

In most cases the hosted version of Voyant is probably still best in terms of convenience and sharing of work, but there are times where a local instance of Voyant may be preferable, especially with respect to performance, privacy and off-line access. This workshop, focused on using (not managing) My Voyant locally, will serve to expand the possibilities of doing digital text analysis.

Workshop Leaders:

Stéfan Sinclair, sgsinclair@gmail.com, is an Associate Professor in Digital Humanities at McGill University. His research focuses primarily on the design, development and theorization of tools for the digital humanities, especially for text analysis and visualization. He has led or contributed significantly to projects such as Voyant Tools, the Text Analysis Portal for Research (TAPoR), and BonPatron. Other professional activities include serving as associate editor of Digital Humanities Quarterly, as well as serving on the executive boards of CSDH/SCHN, ACH, ADHO, and centerNET.

Geoffrey Rockwell, grockwel@ualberta.ca, is a Professor of Philosophy and Humanities Computing at the University of Alberta, Canada. He has published and presented papers in the area of philosophical dialogue, textual visualization and analysis, humanities computing, instructional technology, computer games and multimedia. He was the project leader for the CFI (Canada Foundation for Innovation) funded project TAPoR, a Text Analysis Portal for Research, which has developed a text tool portal for researchers who work with electronic texts. He is the author of "Defining Dialogue: From Socrates to the Internet with Humanity Books."

Target Audience: a wide range of DH practitioners interested in text analysis, particularly for research, teaching, or technical support. Voyant Tools workshops are typically fully-subscribed; we prefer to limit registration to about 30 people to allow us to help participants as needed.

Introduction to electronic books and EPub 3.0

Sperberg-McQueen, Michael

Black Mesa Technologies LLC, United States of America

1. Description

An essential property of modern reading and writing was recognized decades ago by Ted Nelson: we can now read using devices which can evaluate conditional expressions and do different things depending on the result. The full implications of the use of electronic devices as technologies for reading won't be clear for a long while, if ever, but some of them are beginning to be clearer as the history of text presentation with computers continues.

The availability of electronic reading devices touches on the concerns of digital humanities in several ways. As students and analysts of human cultural consumption, digital humanists will study with interest any shift in technologies of reading, even ones whose effects are less seismic in scope than those of

the shift to electronic text. As scholars who deploy the tools of information technology to study texts of the past more effectively, digital humanists should know the capabilities and limitations of current ebook technologies, as they compare with other ways of studying texts. And as producers and consumers of texts for ourselves and for our students, digital humanists will often wish to view ebooks from a producer's or publisher's point of view.

This tutorial offers an introduction to current standards for electronic books, focusing on EPub 3.0 (a standard issued by the International Digital Publishing Federation) with side glances at other specifications. The competition to EPub 3.0 includes proprietary ebook formats and page-image formats like PDF and DejaVu; the supporting specifications for EPub 3.0 include XHTML, HTML 5, CSS, SVG, and Zip.

Many of the challenges of ebook production will be familiar to anyone with experience in book or journal production: mathematics, tables, graphics, and figures are no easier to handle (but, happily, also not much harder) in electronic books than they are in print publications or on the Web. Others will be familiar from web-site production: incompatibilities among ebook readers resemble the incompatibilities among Web browsers in the mid- to late 1990s, and the relation of hardware or software behavior to the prescriptions of the specification remains (to put it gently) complex. Earlier versions of the EPub standard put strict limits on the use of interactive elements; this lowered the threshold for ebook production, but also the ceiling of what was possible in a standards-compliant ebook. So the tutorial will devote special attention to clarifying what is (and is not) made possible by the broader rules of EPub 3.0.

Prerequisites: no firm prerequisites. Participants with some familiarity with XML, HTML, HTML 5, and CSS will be in a better position to follow the details of examples.

2. Contact information

C. M. Sperberg-McQueen
 Black Mesa Technologies, LLC
 259 State Road 399
 Española NM 87532-3170
 Tel. +1 (505) 747-4224 (w) 692-7019 (m)
 Email: cmsmcq@blackmesatech.com

I am an information-technology consultant specializing in problems of preservation and access for cultural heritage materials, publishing systems, and scholarly and public information. My research interests are centered around problems of information modeling and document processing; my practical expertise centers around XML and related technologies.

3. Target audience and expected number of participants

The target audience consists of digital humanists interested in understanding the current generation of open specifications for electronic books and how to exploit them to solve design problems in ebook production.

I've never taught this tutorial before, so I have no relevant experience on which to base an estimate of audience size. Any audience size from 5 to 50 seems possible.

4. Technical support

No requirements beyond the LCD projector, screen, and wifi mentioned on the web site.

5. Outline

A tentative half-day schedule is:

Session 1 (90 minutes)

- Introduction to the course.

- Varieties of electronic books and electronic reading devices; an informal survey (30 minutes).
- Overview of the EPub 3.0 specification and foundational specs. What is specified, where? What is left to vary among products? Role of XHTML, HTML 5, CSS, and Zip in EPub 3.0. (30 minutes).
- Examples: building an EPub 3.0 publication by hand, with XHTML editor, file system, and zip. Tools for automating construction of ebooks. (30 minutes).

Session 2 (90 minutes)

Quick survey of some alternatives to EPub 3.0: EPub 2.0, Kindle formats, PDF, proprietary apps, Web pages.

Problems in ebook delivery:

- gaps in CSS support
- gaps in HTML 5 support
- variations in interpretation of spec
- memory management

Challenges in ebook design:

- mathematics and other formula languages
- graphics, diagrams
- maps
- tables
- multilingual texts
- punctual annotations
- running commentary
- dictionary lookup and other reader aids
- animations
- hyperlinking
- user interaction
- openness to the Web

Concluding words. Where to go from here? Further resources.

A full-day version of the tutorial would cover essentially the same topics, in more detail, and with some opportunities for hands-on work by the participants. I would have a mild preference for a full-day version, but I would also be happy to teach a half-day tutorial.

Using the PressForward Plugin to Create and Maintain Web Publications

Westcott, Stephanie

George Mason University, Roy Rosenzweig Center for History and New Media, United States of America

Fragaszy Troyano, Joan

George Mason University, Roy Rosenzweig Center for History and New Media, United States of America

Abstract:

With the PressForward plugin, the power to publish a curated site that highlights work from the open web is available to everyone. A WordPress plugin that enables users to aggregate and transform web feeds into a site that republishes blog posts, news, and reports, the PressForward plugin streamlines the process of creating web publications. PressForward publications build communities, direct attention to often-overlooked work, and stimulate discussion of ideas, methods, and news. In this half-day workshop, participants will learn to use the PressForward plugin to create their own publications, track workflow through aggregation, review, and nomination, and to publish content from select RSS feeds.

Developed by the PressForward Project at the Roy Rosenzweig Center for History and New Media, the PressForward plugin is the heart of the RRCHNM publication *Digital Humanities Now*. In addition, the publications *Global Perspectives on Digital Humanities*, *Dh+Lib* and the forthcoming

Environmental Humanities Now all use the plugin to maintain a community-driven website that offers readers and participants an opportunity to engage in relevant conversations about their field on the open web. The plugin was developed with an eye toward eventual dissemination and replication, and the timing of DH2014 corresponds with the public release of the plugin and the availability of its documentation.

To get the full benefit of the workshop, participants should bring a laptop, create a WordPress site (directions will be provided prior to arrival), and collect five to ten RSS feeds to begin populating their site. Groups intending to collaborate on a community-driven publication are particularly welcome, and discussion of how to structure and organize a community publication will be included in the schedule.

Intended Length and Format of the Workshop:

8:30 – 9:00: Introduction of class and participants
 9:00 – 9:30: Discussion of how the plugin functions for *Digital Humanities Now*
 9:30 – 9:45: Break
 9:45 – 10:30: Installing the plugin and adding feeds
 10:30 – 11:15: Using the plugin to publish/getting your site up and running
 11:15 – 12:00: Discussion of work flow for a group publication and other uses of the plugin.

Description of Target Audience:

This workshop is open to up to 20 participants of all skill levels with an interest in developing websites that include a component of aggregating and curating work found on the open web. This could include a community of DHers who want to create a publication like *Digital Humanities Now* to highlight the work they believe deserves wider dissemination. Users could also include grad students using the plugin to highlight work on a single topic of interest, a scholar who would like to compile the work they themselves have done elsewhere onto a single site, or professors who may find the plugin a convenient way to highlight work for their students to read on a course site. Given the flexibility of the plugin, the target audience is diverse. There is no need for an advanced call for papers.

Full Contact Information of workshop leaders:

Stephanie Westcott Roy Rosenzweig Center for History and New Media George Mason University, MSN 1E7 Fairfax, VA 22030 703-993-9277 westcott.chnm@gmail.com

Stephanie Westcott is Research Assistant Professor at the Roy Rosenzweig Center for History and New Media at George Mason University. As a member of the PressForward team, she is Managing Editor of *Digital Humanities Now* and *Journal of Digital Humanities* and researches scholarly communication on the open web. A cultural historian with expertise in the history of gender and sexuality in the twentieth-century United States, she received her PhD from the University of Wisconsin-Madison in 2012.

Joan Fraguas Troyano Roy Rosenzweig Center for History and New Media George Mason University, MSN 1E7 Fairfax, VA 22030 703-993-9277 joanfraguas@georgetown.edu

Joan Fraguas Troyano is a Research Assistant Professor and Director of the PressForward project at the Roy Rosenzweig Center for History and New Media at George Mason University. With PressForward she is researching the sourcing, evaluating, publishing, and crediting of scholarly communication from the open web. She also edits two experimental publications — *Digital Humanities Now* and the *Journal of Digital Humanities* — and oversees the development of the PressForward plugin to facilitate the aggregation, curation, and dissemination of scholarship. Joan also is a practicing and teaching public historian with experience

working on the September 11 Digital Archive and Echo projects at RRCHNM, as well as museum exhibition research and education at the Smithsonian National Museum of American History and National Portrait Gallery. At Indiana University she studied music performance and earned a BA in History and Latin. Her PhD is in American Studies from George Washington University, where she researched immigration history, visual culture, and public understandings of the past.

Using CLARIN for Digital Research

Wynne, Martin

martin.wynne@oucs.ox.ac.uk
 Oxford University, United Kingdom

Trippel, Thorsten

thorsten.trippel@uni-tuebingen.de
 Eberhard-Karls-Universität Tübingen

Draxler, Christoph

draxler@phonetik.uni-muenchen.de
 Ludwig-Maximilians-Universität München

Short description

Many researchers in the digital humanities will have heard of CLARIN, and the efforts to build a persistent, reliable and sustainable infrastructure for language resources and tools. Since the first plans in 2008 CLARIN has been under way, first in a preparatory phase project to prepare the way for funding and to set up the organizational structures, and since 2011 in the construction phase to build infrastructure services and integrate resources. CLARIN aims to provide services to support and facilitate support the use of digital language resources and tools in the humanities and social sciences, and has been adopted as a key service in the national roadmaps for research infrastructure in numerous European countries. Now is a good time to ask "What can CLARIN do for me?" and "What can I do for CLARIN?". This tutorial workshop will aim to give a practical introduction to CLARIN ,focussing on providing answers to these key questions.

The workshop will include following topics:

1. How can I find resources using CLARIN?
 - Locating resources using the Virtual Language Observatory (VLO)
 - Searching for and in resources with Federated Content Search
 - Accessing resources via the PID
 2. How can CLARIN archive and curate my resources?
 - Depositing services
 - Identifying the right archive
 - Recommendations on the creation and archiving of resources from the CLARIN community
 3. How can I use CLARIN to make more impact with my research?
 - Citing data using PIDs (Citation recommendations)
 - Open access policy to data (reuse of research data)
 4. How can I integrate my tools and services with the CLARIN infrastructure?
 - making a repository known to CLARIN
 - making a service known to CLARIN
 - integrating tools as web services in existing environments
 5. How can I do research with CLARIN?
 - showcases of exemplary collaborative research in the Humanities
 - opportunities for future collaborations
- The anticipated audience will fall into two categories (although a significant number of individuals will play both roles): (i) creators, developers and curators of language resources and tools, and (ii) researchers in the humanities

and social sciences who are users or potential users of these resources and tools. The workshop hopes to offer something to both categories of people, and to gain extra benefits to all by bringing them together, and allowing users to understand how services are built and maintained, and for developers to understand more about how researchers use services.

Among the outcomes will be a series of accessible guides, or HOW-TOs, online on the CLARIN portal, and maintained and updated by CLARIN staff. Organizers and speakers will attend with funding from national CLARIN initiatives. The CLARIN European coordinating office will offer bursaries for early career and postgraduate researchers to attend the workshop.

The session will be half a day, with 3 hours of sessions plus breaks. A suggested timetable is given below.

09:00	<i>registration</i>
09:30	Welcome and introductions
09:45	How can I find resources using CLARIN?
10:30	<i>break</i>
11:00	How can CLARIN archive and curate my resources?
11:30	How can I use CLARIN to make more impact with my research?
11:45	How can I integrate my tools and services with the CLARIN infrastructure?
12:00	How can I do research in the humanities with CLARIN?
13:00	<i>end</i>

Organisers

Martin Wynne is Director for User Involvement for the CLARIN European Research Infrastructure, and a senior research support officer and digital research specialist at the University of Oxford. He has worked in corpus linguistics and related areas for more than twenty years, and has played a leading role in a number of support and infrastructure services such as the Oxford Text Archive, Arts and Humanities Data Service, and Project Bamboo, and he was the originator of the Digital Humanities at Oxford initiative.

martin.wynne@it.ox.ac.uk

Thorsten Trippel works in the area of language resource management and is a specialist on metadata for language resources, representation models and infrastructures. His expertise also covers lexical resources and terminology, text technology and repository systems. As a national expert in standardization of language resources he works on Persistent Identification of Language Resources, metadata formats and data categories, both in national and international standardization settings. Within CLARIN he is involved in operations on the European level and is one of the coordinators of the German CLARIN-D, especially focusing on the requirements from the Humanities and Social Sciences for the research infrastructure.

thorsten.trippel@uni-tuebingen.de

Christoph Draxler is the representative for spoken language resources in CLARIN-D. He is located at the Bavarian Archive for Speech Signals at Ludwig Maximilian University Munich. He has collaborated in many large-scale speech data collection projects such as SpeechDat, Ph@tSessionz, or VOYS. His research interests are speech databases and web-based tools for speech recording, annotation, online experiments and speech-related crowdsourcing. draxler@phonetik.uni-muenchen.de

Tutors and presenters will be selected from the CLARIN community and networks on the basis of their expertise and presentation skills. Bursaries to cover the costs of participation will be available from CLARIN.

Curation, Management, and Analysis of Highly Connected Data in the Humanities

de la Rosa, Javier

University of Western Ontario, Canada

Brown, David Michael

University of Western Ontario, Canada

This half-day workshop will instruct participants in the use of SylvaDB [1] to manage large sets of highly connected and semantically rich data. Beginning with raw metadata gleaned from cultural objects, participants will learn how to design a productive data model, store the data according to that model, and administer it collaboratively. Furthermore, they will learn how to integrate the data with other applications, analyze it using a powerful analysis framework, and organize their results logical collections called reports. The following proposal outlines the workshop and its relevance in four sections: 1) an introduction with an overview of the SylvaDB database management system and its applicability in a humanities context, 2) expected participant outcomes, 3) workshop content, and 4) a brief conclusion.

1. Introduction

1.1 SylvaDB

SylvaDB is a browser based database management application developed by the CulturePlex Lab at Western University Canada. Written on top of the Neo4j graph database backend, SylvaDB allows users to create their own databases, each with an easy to use interface for: designing flexible data models, performing Create, Read, Update, and Delete (CRUD) operations, controlling user permissions, building and executing graph style queries, analyzing/visualizing data. These features were specifically designed to empower non-programmers by providing them user-friendly access to the power and flexibility of the graph database data structure.

For developers, SylvaDB features a streaming API that facilitates integration with new or existing applications. The API is implemented as a RESTful service that supports input and output for read/write operations and data analysis procedures. Furthermore, SylvaDB features a set of graph algorithms that can be run by users in server-mode. Finally, SylvaDB is an open source project, which, simply put, promotes transparency and customizability by allowing developers to fully understand the product they are using.

1.2 SylvaDB in the Humanities

SylvaDB was originally designed to handle the problems of data storage, management, and analysis specifically encountered in a Digital Humanities research context. The advent of innovative and increasingly sophisticated methods for analysis in the humanities results in an increased need for computational infrastructure to store and process data [2]. However, access to this infrastructure is not necessarily equal, and many non-programmers lack the technical expertise to implement the necessary solutions, or the resources to hire a programmer to do it. SylvaDB overcomes this limitation by providing a powerful, easy to use framework for data management and processing.

Part of SylvaDB's power rests in the technology upon which it was built: Neo4j's graph database. This model for data storage provides the base for a system that is at once powerful, semantically rich, and flexible. These characteristics directly corresponds to challenges presented by humanities research:

- Humanities data can be messy, unsure, or likely to change at some point, hence necessitating a flexible storage framework [3]. SylvaDB provides an interface for the user to design a

flexible data model that best fits their data, and change it as necessary.

- In the case of highly interconnected—or network—data [4], analysis can be quite costly in terms of both time and memory. SylvaDB supports native graph style queries designed for traversing millions of nodes and relationships in milliseconds and providing an efficient way to generate network and descriptive statistics. It is then easy to analyze query results using SylvaDB’s flexible data analysis environment to produce rich, interactive visualizations.
- Humanities data is semantically rich, often taking the form of highly structured and interconnected metadata, which is difficult or inefficient to manage using SQL database technology. SylvaDB utilizes graph database models to allow semantic information stored as types and attributes in both nodes (data points) and the relationships between them, effectively facilitating semantic querying capabilities.

2. Outcomes

This workshop focuses on SylvaDB as a tool that empowers non-programmers to take control of their data; however, in a broader sense, its goal is to explore the concepts and practices behind effective data storage, management, and analysis. The specific learning outcomes for the workshop are as follows:

- Mastery of the entire SylvaDB application package including: data modeling, CRUD operations, data administration, permissions controls, data import/export, query building, analysis, and report generation.
- Awareness of fundamental database data modeling concepts: objects, types, attributes, relationships, schemas.
- Practice with design thinking for data modeling—designing your model to solve a specific humanities problem.
- Awareness of different database storage models and their pros and cons in a humanities context.
- Expanded knowledge of types of data analysis/visualization, the reasoning behind them, as well as their usefulness to better understand complex humanities problems.

3. Content

The content of the workshop will be presented in three sections: 1) a general overview of databases and the associated concepts, 2) a database building activity that introduces SylvaDB and its features, and 3) an experiential learning session in which small groups model, store, and analyze a real data set. Participants are strongly encouraged to create accounts at testing.sylvadb.com prior to the workshop, and bring their laptops.

3.1 Overview

The goal of this section is to introduce essential concepts and terminology associated with databases. Beginning with the concept of data types and attributes, participants will be exposed to different models for data storage: relational tables, document/key-value stores, and graphs. Real world examples will be provided of each type of database, along with a discussion of the potential use cases and advantages/disadvantages of each system. Here the focus will fall primarily on the motivation behind using each storage method, and provide an introduction to the concepts behind data modelling.

3.2 SylvaDB Use and Features

After the participants are familiar with the basics of data storage, we will see how these concepts have been applied in SylvaDB. The instructors will provide a quick introduction of the SylvaDB software package and its features, focusing particularly on data model (schema) creation, CRUD operations, building queries, and visualizing query results. This will be presented as a live demo using SylvaDB, and participants will be encouraged to follow along using their own

laptops. Next, the participants will be presented with a small, easy to model data set. As a class, we will learn how to build and process a graph database, encouraging participation and student input regarding the following processes:

- Schema creation: Students will learn to utilize the full capabilities of SylvaDB’s schema creation interface. This section will exemplify the process of creating schema types to represent different types of data, adding attributes to the types, determining relationships between them, and adding semantic annotation to the relationships. In this section, instructors will emphasize the importance of purposeful schema creation in order to produce insightful results during the analysis phase.
- Data management: Students will learn to store and manage data using the schema we have created. The participants will become familiar with performing CRUD operations and controlling collaborator permissions. Also, this section will include an overview of using SylvaDB’s tabular data display to search the database, with emphasis on how to use the built-in filters for maximum search efficiency.
- Analysis: The instructors will present an expanded dataset based on the previous example schema and data. This data will be used to familiarize the participants with SylvaDB’s query builder and data analysis environment. Participants will learn how to build several different type of queries and visualize their results using the built in data analysis environment. During this process, instructors will focus on providing examples of querying and visualization practices that fit the unique characteristics of the data set.

3.3 Small Group Activity

Participants will use what they have learned to model, store, and analyze a real humanities data set. The instructors will present a data set that consists of metadata gleaned from library holdings. In small groups, participants will evaluate the data and determine how it could best be modelled using SylvaDB. After a brief discussion of possible data models, we will build a standard schema model that will allow each group to import pre-configured data into their database. This reduces the complication and time commitment of manually inputting data, and allows the groups to focus on building effective queries and visualizing the results. Each group will design a series queries to visualize whatever aspect of data they choose, and then configure a report that includes their preferred visualization. At the end of the workshop, drawing from the results obtained, students will present and discuss their mini-project conclusions.

4. Conclusion

This workshop presents SylvaDB as a tool that enables non-programmers to harness the full power of a graph database, create expressive and flexible data models, and perform complex analytical procedures. Upon completing this workshop, participants will not only have learned to use a powerful software package, but also the fundamental concepts behind databases, data modeling, and analytics. Perhaps most importantly, this knowledge will inspire confidence in humanities practitioners that move in a field increasingly focused on data [5], enabling them to take their research to new heights and levels of excellence.

5. References

1. de la Rosa, J., Suárez, J.L., Sancho, F. *SylvaDB: a Polyglot and Multi-Backend Graph Database Management System*. DATA Conference, Iceland. 2013.
2. Poole, Alex H. "Now Is the Future Now? The Urgency of Digital Curation in the Digital Humanities." 7.2 (2013): n. pag. Digital Humanities Quarterly. Web. 10 Feb. 2014.
3. Schöch, Christof. "Big? Smart? Clean? Messy? Data in the Humanities." Journal of Digital Humanities. N. p., 22 Nov. 2013. Web. 10 Feb. 2014.

4. **Meeks, Elijah.** "Modeling Transportation in the Roman World: Implications for World Systems." *Leonardo* 46.3 (2013): 278. Print.
5. **Fallon, Dorothy.** "Big Data in the Humanities: The Need for Big Questions." *Science in Culture*. Web. 14 Feb. 2014.

6. Outline

1. Intro to database concepts - 20 min
 1. Data types, attributes, and relationships
 2. Models for storage and their pros/cons
 1. Relational
 2. Key-value/document
 3. Graph
2. General overview of SylvaDB's features - 20 min
 1. Schema Creation
 2. CRUD operations
 3. Searches/Filtering
 4. Query Builder
 5. Analysis Environment
3. Activity 1 - Group Database Building Activity - 50 min
 1. Data modeling and Schema Generation
 2. Data Entry
 3. Management - User Permissions Searches
 4. Analysis - Query Building Activity
4. Activity 2 - Small Group Data Modeling and Analysis - 60 min
5. Presentation of Results and Discussion - 30 min

Workshop Leaders

Javier de la Rosa

versae@gmail.com
CulturePlex Lab. University of Western Ontario
519-661-2111 Ext. 89251

Javier is a 3rd year PhD student at Western University. His general research interests are in graphs, graph databases, query languages, complex networks, and temporal ontologies. His main research interest is in Network Theory.

David Brown

dbrow52@uwo.ca
davidmichaelbrown1@gmail.com
CulturePlex Lab. University of Western Ontario
519-661-2111 Ext. 89251

David is a 2nd year Ph.D. student at Western University. His primary research interests are: graph databases, network analysis, maps, Mesoamerican culture, New Spain, and web development using Python and Javascript. He is currently developing his expertise in applying data-intensive analysis techniques to shed light on questions from the humanities and social sciences.

Elika Ortega

eortegag@uwo.ca
CulturePlex Lab. University of Western Ontario
519-661-2111 Ext. 82822

Elika is a Postdoctoral Fellow at the CulturePlex Lab, Western University. Her research focuses on narrative in digital media and the study of narrative networks. She is especially interested in the ways in which digital media have revitalized the sociality of narrative and the interactions of print and digital media, as well as on the network structures of convergence media texts.

Juan Luis Suarez

jsuarez@uwo.ca
CulturePlex Lab. University of Western Ontario
519-661-2111 Ext. 85858

Juan Luis is a Professor of Hispanic Studies in the Modern Languages and Literatures Department as well as the Director of the CulturePlex Lab at Western U. His research deals with cultural complexity and complexity theory, digital humanities, technologies of humanism, Hispanic Baroque, as well as globalization and new literatures. Some of his books are "Tecnologías del Humanismo", "Herederos de Proteo", and "Calderón: El escenario de la imaginación". Very recently, he

also spearheaded a successful IDI proposal at Western U. in the field of Digital Humanities on which he is collaborating with participants from a broad spectrum of fields of study.

Target Audience

This workshop is intended for anybody interested in learning the skills to better model, store, manage, and analyze data. It is of particular interest to researchers that deal with highly connected data, and are interested in harnessing the power of the graph database for storage and analysis. No programming skills are necessary, and no previous knowledge of databases is required; however, the focus on graph databases and the SylvaDB toolkit makes this workshop relevant for experienced database users and developers. In past, much smaller conferences, our SylvaDB workshop has attracted approximately 20-30 people.

Panels

Annotating in Digital Music Edition - concepts, processes and visualisation of annotations

Beer, Nikolaos

nikolaos.beer@uni-paderborn.de

Universität Paderborn, Musikwissenschaftliches Seminar Detmold/
Paderborn

Bohl, Benjamin W.

bohl@edirom.de

Universität Paderborn, Musikwissenschaftliches Seminar Detmold/
Paderborn

Seuffert, Janette

seuffert@em.uni-frankfurt.de

Goethe-Universität Frankfurt am Main, Institut für Musikwissenschaft

Annotating in Digital Music Edition - concepts, processes and visualisation of annotations

1. Organization

- Nikolaos Beer
- Benjamin W. Bohl
- Janette Seuffert

2. Confirmed speakers

- Nikolaos Beer, Dariah-DE, Universität Paderborn
- Stefanie Steiner-Grage and Frank Zalkow, Reger-Werkausgabe, Max-Reger-Institut, Karlsruhe
- Janette Seuffert, OPERA, Goethe-Universität Frankfurt am Main
- Benjamin W. Bohl, Freischütz Digital, Universität Paderborn
- Christine Siegert and Kristin Herold, A Cosmopolitan Composer in Pre-Revolutionary Europe – Giuseppe Sarti, Universität der Künste Berlin

3. Overview

This panel presents theoretical and technical approaches in digital scholarly music editions focusing on different forms of annotations, data enrichment and data visualization.

Annotating is one of the central techniques in humanities to gather, analyse, discuss and present information about works, sources and their interpretation. In traditional scholarly music edition this means to create extensive catalogues of verbal critical comments on findings (critical apparatus) in one or more corresponding music and text sources. The critical report in print editions, containing the critical apparatus and also general further information about the work and its sources, is usually stored separately from the music editions. Considering the discrepancy between the verbal description of a finding and its representation in notated music, using such linear structured catalogues often requires an intensive effort to read.

The introduction of digital techniques in recent scholarly music editions not only allows an extension of the amount of considered material (editions, digitized images of music and text sources, contextual information), but also facilitates the presentation and immediate contextualization of the edition and its sources within the critical report¹.

At first this can be realised as a transformation of traditional critical commentary into a digital form by visualizing it on and linking it to digital representations (images) of editions and sources. With current developments of specialized encoding formats, more elaborate concepts of annotations come into consideration, moving away from commenting on music and text verbally to an encoding of variants, readings or interrelations with the help of markup techniques².

The Music Encoding Initiative (MEI)³ as an XML-based markup standard provides possibilities to encode data of musical works such as the content itself (music notation and text), bibliographical metadata, structural data of a work and its sources (for instance acts, scenes, sections, parts, bars), and data on corresponding representations like performances. This opens possibilities to embed different types of annotations and/or to link MEI-data to other structured data like texts encoded according to the Guidelines of the Text Encoding Initiative (TEI)⁴

Over the last decade, various research and music editing projects have been exploring the application of digital technologies for their purposes, and conclusively it can be said that "There is no doubt that digital media already have influenced, and in the future will in some ways fundamentally change our conception of modern editions and editorial practices"⁵.

For example the Edirom-Project⁶ at the Musikwissenschaftliches Seminar Detmold/Paderborn (Musicology Seminar Detmold /Paderborn) has developed a set of software tools, the so-called Edirom Tools, to support music edition projects in all work stages from data compilation (Edirom Editor) to publication (Edirom Viewer or Edirom Online)⁷. Four edition projects currently using the Edirom Tools and additional technologies, will present themselves in the following, covering editions of musical works from different eras and multiple genres. They share in common several approaches to examine and analyse opportunities and changes in the edition process derived from the application of digital techniques. Exploring the process and an extended concept of annotating is thereby of particular interest.

4. Presentations

4.1 Understanding and encoding musical variants and readings as annotations

As an introduction to this panel, MEI principles for inline data enrichments will be presented, allowing to encode information on variants and readings as musical text. It will be also discussed in which way such encodings could be understood as annotations and how they could be visualized in music notation software tools for reuse in different contexts of musicological research.

The MEI Score Editor (MEISE)⁸, developed as part of the infrastructure project Dariah-DE⁹ and designed to facilitate editing and visualization of MEI encodings and data enrichments, will be presented as an example application. Considering one of Dariah's main research topics – techniques of scholarly annotating and annotation – MEISE's key feature is the possibility to handle encoding and visualization of variants in MEI. MEISE will enable edition projects to maintain digital research methods.

4.2 Hybrid Edition Max Reger-Werkausgabe

One of the first projects using digital techniques and the Edirom Tools is the Hybrid Edition *Max Reger-Werkausgabe* (RWA)¹⁰ at the Max-Reger-Institute, Karlsruhe, which is published both as printed and digital edition. Focusing on the organ works in the first module, the two striking features of the edition's digital part are the prioritised linked-in annotations and its complex encyclopedia as an extensive collection of references to Max Reger's life and œuvre¹¹.

Several considerations had to be made about organizing the collaborative edition process and the two-way publication model, like centralised capture of annotations in a database for further semi-automated publication processing. This allows for the editor to postpone publication based decisions to later stages in the edition process.

Already when entering the annotations into the database, the editor assigns a priority level according to their respective importance. The lowest priority (= priority level 3) comprises

remarks which are relevant only because of philological reasons (such as differing warning accidentals in the various musical sources), while the higher categories (= priority levels 1 and 2) concentrate on important matters concerning either the works themselves or their tonal realization. According to these priorities the remarks are guided either into both the digital and printed editions (levels 1 and 2) or, if of lower importance (level 3), only into the digital version.

Similar collaborative solutions as for the annotations had to be found for the preparation of the digital encyclopedia, containing texts (encoded in TEI) and images, to allow for linking to and between all edition parts and contents.

4.3 OPERA – Spektrum des europäischen Musiktheaters in Einzeleditionen

In the *OPERA*¹² research and edition project, 21 musical dramatic compositions, drawn from different periods and different genres such as opera, Singspiel, drama with incidental music, melodrama, or ballet are being edited. A collaborative team of internal and external editors publishes its editions in a hybrid form: the score is published in traditional book format; the critical report, including images of the Music Edition and sources, and the Text Edition are presented in an electronic Edirom-based form.

The Edirom software has been modified in order to integrate the TEI-based Text Edition into the Edirom environment and to interlink it with other digital resources (digitized score and source images) through synchronization of specific structural units/data of music and text (bars, lines, numbers, etc.). The modification is compatible with future tasks such as the integration of audio and video resources. Both, music and text edition share one common critical apparatus, whose annotations are strictly categorized as Music (M), Text (T), and Stage (S), and reflect the peculiarities of musical theatre¹³. In addition to the annotations, which are mostly limited to report variant readings, OPERA uses a comment window for historical informations and contextualization.

4.4 Freischütz Digital – Paradigmatische Umsetzung eines genuin digitalen Editionskonzepts

Inspired by Wiering's multidimensional Model¹⁴, the approach of *Freischütz Digital*¹⁵ is a genuine digital edition of the opera *Der Freischütz* by Carl Maria von Weber, considering all current theoretical and technical possibilities for the edition process. The principal part of this edition is the encoding and interlinking of music sources, libretto sources, and further material. The transcriptions of all music texts are encoded according to MEI whereas the transcriptions of the libretto texts and their stages of development are encoded according to TEI. For the first time, acoustic elements (several sound files of recorded performances), are analysed and visualized in the digital edition in a way that they are synchronized with and set in context to the music and text sources, their transcriptions and annotations¹⁶. This opens new possibilities for processing, enriching and reusing the edition's data.

Annotations for the music sources will be based on the print edition currently under preparation at the Carl-Maria-von-Weber-Gesamtausgabe (WeGA)¹⁷ and be transferred to XML-based encoding. For the purpose of annotating the libretto sources the project develops a stand-off encoding model and a corresponding tool (the so-called CoreBuilder) facilitating the generation of associations¹⁸.

4.5 A Cosmopolitan Composer in Pre-Revolutionary Europe – Giuseppe Sarti

In the research project *A Cosmopolitan Composer in Pre-Revolutionary Europe – Giuseppe Sarti*¹⁹ two MEI based Edirom editions of Sarti's operas *Fra i due litiganti il terzo*

gode and Giulio Sabino are prepared, aiming to consider appropriately the fact that Italian opera as a genre does not intend to provide stable works with an "authentic" text. Instead, the operas were always adapted to the conditions of new performances.

Our edition presents the sources and their readings in a non-hierarchical way, with the intention to show a wide range of variants and arrangements²⁰. Therefore, we distinguish between different types of annotations: We do not only document pure errors in the musical text, but also explain the textual variants and establish – according to the FRBR (Functional Requirements for Bibliographic Records) model²¹ – the relationships of the sources. This approach seems to have clear advantages compared to the typical entries in common critical reports.

References

1. Joachim Veit (2010): *Es bleibt nichts, wie es war – Wechselwirkungen zwischen digitalen und 'analogen' Editionen*. In: editio, vol. 24, Berlin/Boston, pp. 37–52.
2. Johannes Kepper (2011): *Musikdition im Zeichen neuer Medien – Historische Entwicklung und gegenwärtige Perspektiven musikalischer Gesamtausgaben*, Norderstedt.
3. The Music Encoding Initiative (MEI) provides extensive guidelines and corresponding schemata for XML-based encoding of music. Its special features for academic and scholarly purpose make it unique among other encoding formats for music. Available online at www.music-encoding.org (last accessed: March 6, 2014).
4. The Text Encoding Initiative (TEI) maintains comprehensive guidelines and corresponding schemata for XML-based text encoding, available online at www.tei-c.org (last accessed: March 6, 2014).
5. Bjarke Moe, Axel Teich Geertinger (2008): *Digital Editions of Music. Perspectives for Editors and Users*. Proceedings, p. 7. Online at: www.bjarkemoe.dk/digitaleditions2008.pdf (last accessed: March 6, 2014).
6. See the project's website at www.edirom.de (last accessed: March 6, 2014).
7. Benjamin Bohl, Johannes Kepper, Daniel Röwenstrunk (2011): *Perspektiven digitaler Musikditionen aus der Sicht des Edirom-Projekts*. In: Die Tonkunst 5, pp. 270–276.
8. See the project's website at de.dariah.eu/web/guest/mei-score-editor (last accessed: March 6, 2014).
9. See the online portal of DARIAH-DE at de.dariah.eu (last accessed: March 6, 2014).
10. See the project's website at www.max-reger-institut.de/de/rwa.php (last accessed: March 6, 2014).
11. Alexander Becker, Christopher Grafschmidt, Stefan König, Stefanie Steiner (2013): *Möglichkeiten und Konsequenzen der Digitalen Musikdition am Beispiel der Reger-Werkausgabe (RWA)*. In: Medienwandel/Medienwechsel in der Editionswissenschaft, ed. Anne Bohnenkamp (= Beihefte zu editio 35), Berlin/Boston, pp. 159–166.
12. See the project's website at www.opera.adwmainz.de (last accessed: March 6, 2014).
13. Janine Droeße, Norbert Dubowy, Andreas Münzmay, Janette Seuffert (2013): *Musik – Theater – Text. Grundfragen der Musiktheaterphilologie im Spiegel der OPERA-Hybridausgaben*. In: editio, vol. 27, Berlin/Boston, pp. 72–95.
14. Frans Wiering (2009): *Digital Critical Editions of Music: A multidimensional Model*. In: Gibson Crawford: Modern Methods for Musicology, Ashgate, pp. 23–45.
15. See the project's website at www.freischuetz-digital.de (last accessed: March 6, 2014).
16. Meinard Müller, Thomas Prätzlich, Benjamin Bohl, Joachim Veit (2013): *Freischütz Digital: A Multimodal Scenario for Informed Music Processing* (conference paper), WIAMIS 2013 – 14th International Workshop on Image and Audio Analysis for Multimedia Interactive Services, Paris, July 3.
17. For further details on the Carl-Maria-von-Weber-Gesamtausgabe see the project's website at www.webergesamtausgabe.de (last accessed: March 6, 2014).
18. Raffaele Viglianti, Solveig Schreiter, Benjamin Bohl: *A stand-off critical apparatus for the libretto of Der Freischütz*

(conference paper), TEI Conference and Members Meeting 2013, Rome, October 5, 2013.

19. See the project's website at www.udk-berlin.de/musikwissenschaft/sarti (last accessed: March 6, 2014).

20. For a paradigmatic encoding of one aria see Johannes Kepper, Christine Siegert: Oper multimedial. Zur Edirom-Ausgabe von Haydns Arienbearbeitungen. In: Medienwandel/Medienwechsel in der Editionswissenschaft, ed. Anne Bohnenkamp (= Beihefte zu editio 35), Berlin/Boston 2013, pp. 141-150.

21. **Kristina Richts** (2013): *Die FRBR customization im Datenformat der Music Encoding Initiative (MEI)*, Köln. Online available at publiscolgne.fh-koeln.de/frontdoor/index/index/docId/144 (last accessed: March 6, 2014).

<audio>Digital Humanities</audio>: The Intersections of Sound and Method

Clement, Tanya

tclement@ischool.utexas.edu

School of Information, University of Texas at Austin

Kraus, Kari

College of Information Studies, University of Maryland, College Park

Sayers, Jentery

Maker Lab in the Humanities, University of Victoria

Trettien, Whitney

Soundbox, Franklin Humanities Institute, Duke University

Tcheng, David

Illinois Informatics Institute at the University of Illinois at Urbana-Champaign

Auvil, Loretta

Illinois Informatics Institute at the University of Illinois at Urbana-Champaign

Borries, Tony

Illinois Informatics Institute at the University of Illinois at Urbana-Champaign

Wu, Min

Department of Electrical and Computer Engineering, University of Maryland, College Park

Oard, Doug

College of Information Studies, University of Maryland, College Park

Hajj-Ahmad, Adi

Department of Electrical and Computer Engineering, University of Maryland, College Park

Su, Hui

Department of Electrical and Computer Engineering, University of Maryland, College Park

Lingold, Mary Caton

Soundbox, Franklin Humanities Institute, Duke University

Mueller, Daren

Soundbox, Franklin Humanities Institute, Duke University

Turkel, William J.

The Lab for Humanistic Fabrication, Western University

Elliott, Devon

Western University

<audio>Digital Humanities</audio>: The Intersections of Sound and Method

A wide range of interdisciplinary scholarship on sound has sparked investigations into the cultural histories of aurality and

sound reproduction, the politics of the voice and noise, urban soundscapes, ethnographic modernities, acoustemologies, and the sonic construction of gender, race, and ethnicity.

[i] These important qualitative studies, moreover, have in recent years been supplemented by large-scale quantitative analyses of speech and music datasets, several of which have been underwritten by the International Digging into Data Challenge, including the "Structural Analysis of Large Amounts of Music" (SALAMI) and the "Mining a Year of Speech" projects. Yet a lingering textual bias within digital humanities – largely a product of the field's emergence from textual and literary studies – has obscured the significance of this work for the field, often preventing meaningful overlap. Copyright restrictions, the difficulties of archiving audio formats, and the general lack of tools for researching and writing in audio have contributed to the difficulty of working with sound in digital projects. Aside from the occasional use of CD appendixes or supplementary websites, for example, many studies have not taken full advantage of the affordances of digital media to produce scholarship that integrates audio content into scholarly argumentation. It is against this backdrop that leading sound theorist Jonathan Sterne has argued that "existing digital humanities work has largely reproduced visualist biases in the humanities" (2011).

By identifying and highlighting four research initiatives clustered around audio artifacts, this panel aims to bring sound scholarship and digital humanities into a more meaningful conversation with each other. As these projects demonstrate, sound is materially constituted, containing invisible environmental fingerprints or leaving physical traces in artifacts; and, further, is performative and temporally mediated. Thus to access and analyze sound requires not only a new approach to "tool making" within digital humanities, but a deeper engagement with media studies, archival science, and creative forms of scholarship more generally. As Trettien and Lingold's Soundbox initiative shows, the methodological vibrancy of the field is also predicated on innovation and reform of our critical infrastructures, including the development of publication environments that can take advantage of the cross-medial character of much sound research. Elliott's kits for cultural history, for example, allow users to experience the past through multiple sensory channels, including sight, sound, and touch; and Clement and Kraus's work incorporates extensive spectrographic analysis. Thus a larger aim of the panel is to draw attention to the richly synaesthetic nature of digital sound studies.

Access and Analysis, Tanya Clement (15 minutes)

There are few analysis tools available for humanists interested in accessing and analyzing audio archives that comprise significant artifacts of bygone oral traditions represented in storytelling, speeches, oral histories, and poetry performances. In response to this lack, the iSchool at UT-Austin and the Illinois Informatics Institute (13) at the University of Illinois at Urbana-Champaign (UIUC) hosted a year-long NEH-funded Institute for Advanced Topics in the Humanities called High Performance Sound Technologies for Analysis and Scholarship (HiPSTAS). HiPSTAS included twenty humanities junior and senior faculty and advanced graduate students as well as librarians and archivists interested in analyzing large audio collections. As this speaker will address, HiPSTAS has yielded significant results for audio big data analysis in the humanities including an implementation of the ARLO (Adaptive Recognition with Layered Optimization) software, a machine learning application for analyzing sound on Stampede, an NSF petascale HPC system at the Texas Advanced Computing Center. Originally developed to classify and analyze bird calls by extracting audio features and displaying the audio data as a spectral graph (Downie et al. 2008, Punyasena et al. 2012), ARLO has also been used by humanists as part of HiPSTAS to extract basic prosodic features such as pitch, rhythm and timbre for matching, discovery (clustering) and automated classification (prediction or supervised learning) (Figure 1). This talk will discuss how significant sonic patterns of interest to humanists are discoverable using ARLO with the PennSound

poetry archive and the University of Texas Folklore Center Archives, among other collections.

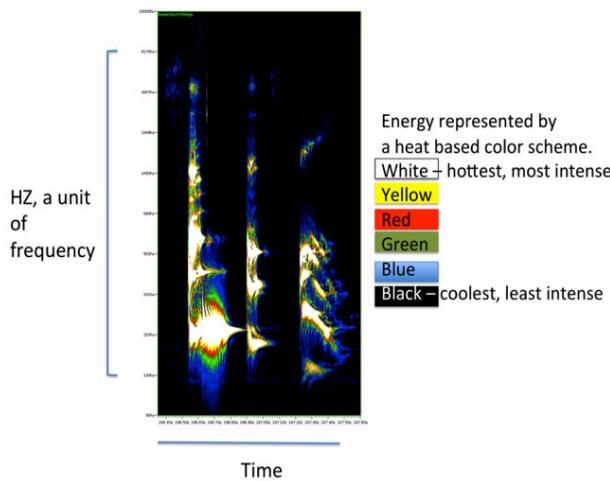


Fig. 1: This spectrogram, created in ARLO, shows Gertrude Stein reading "Some such thing" from her novel *The Making of Americans*; each row of pixels is a frequency band presented across an X-access of time.

Media Archaeology, Devon Elliott (15 minutes)

Recent research in media archaeology (Fuller 2005, Gitelman 2006, Kirschenbaum 2008) underscores why the material particulars of technology matter where questions of culture are concerned. This research is frequently anchored in archival documents—including lab notebooks, patents, and engineering journals—that correspond with technological experiments. Building on this research, this talk shares initial findings from the "Kits for Cultural History" project, a collaboration between the Maker Lab in the Humanities (UVic), the Lab for Humanistic Fabrication (Western), and several memory institutions across Canada. The project involves making physical kits that encourage scholars to reconstruct historical experiments through the use of schematics, facsimiles, and rich media. Audio is central to a number of these kits, especially kits that focus on sound reproduction. Not only does it add another modality to research that is usually text-based or visual in character. It also emphasizes how any media history is a history of the senses: a history of how embodied behaviors like listening relate recursively with technological developments. With audio in mind, the talk argues for the relevance of experimental reconstruction to digital humanities, highlighting the importance of: 1) old technologies to contemporary computing practices, 2) multimodal learning and applied methods to media history, 3) integrating museum collections into these methods, and 4) understanding sound as necessarily material, subject to techniques commonly found in, say, textual studies. These four points draw together domains all too often parsed: visual and sonic paradigms, critical thinking and critical making, media archaeology and digital humanities.

Signal and Noise, Kari Kraus (15 minutes)

Twentieth-century recorded sound, like the first electric power system, originated in Thomas Edison's Menlo Park laboratory shortly before the turn of the century (Hughes, Morton). In the decades that followed, sound technology and power transmission would continue to develop in tandem. In this presentation we introduce an unexpectedly useful consequence of the historic entanglement of sound and electricity: the ability to code our past for time and place. A new collaboration at the University of Maryland aims to recover the date and time on which an historic recording was made based on analysis of incidentally captured traces of small variations in the electric power supply at the time of recording (Oard, et al; Su, et al.). Although the field of audio forensics has used such Electric Network Frequency signatures to authenticate contemporary

recordings for over a decade, our project seeks to extend the period for which baselines are available a further half century into the past. We do this by assembling recordings that were made at known times and comparing their ENF signatures with the signatures in recordings for which we lack such provenance information.

After summarizing the results of our initial experiments, we focus on implications for archival practice, including retention of the original ENF signal across media formats (Figure 1), and conclude on a theoretical note: because ENF is traditionally dismissed as electronic noise by audio engineers and regarded as non-semantic in character, it poses an interesting challenge to the well-established archival concept of "significant properties."

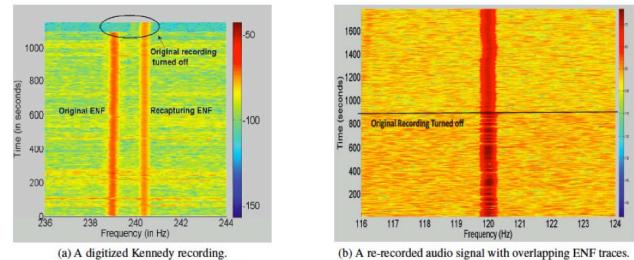


Fig. 2: Analog recordings that have undergone digital conversion and reformatting will often contain two or more ENF signatures: an original and a recaptured signature. The spectrogram in the image on the left shows the ENF trace from a 1962 magnetic tape recording of an oval office meeting during the Kennedy administration and a separate ENF trace embedded at the time of digitization. In the figure on the right, two signatures overlap. We have developed preliminary techniques for distinguishing these multiple traces.

Publication and performance, Whitney Trettien and Mary Caton Lingold (15 minutes)

Soundbox is a collaborative exercise in producing and publishing sonic scholarship. Its main research output is an edited digital collection bringing together a vanguard of emerging scholars and critical artists engaging in sonic scholarship, from exhibits and installations to digital essays, soundscapes, and speculative digital tools.

While the original goal of the project was to show, through example, the wide range of possibilities for an amplified digital humanities, the impossibility of publishing this work through standard scholarly venues – that is, those that facilitate the forms of peer review required for advancement in the profession – has become clear as the project proceeds. Because of concerns over long-term maintenance of digital scholarship, database-driven platforms like Omeka, Scalar, and Wordpress are quickly becoming the standard publishing format for digital work. Designed around arguments written in text and image, though, these platforms are largely inadequate to scholarship that integrates sound beyond the occasional linked audio clip. Thus the potential for amplified scholarly production opened up by, for instance, creative, small-scale, targeted uses of the HTML5 audio tag remains largely unrealizable within an increasingly calcified digital publishing infrastructure – a fact with ongoing consequences for what "counts" as digital humanities scholarship.

Using Soundbox's experience as a case study, the speakers address the structural biases that continue to silence digital humanities. We argue for balancing the need for long-term maintenance and accessibility with a pluralistic approach that does not foreclose the possibilities of new forms and formats.

References

- Attali, Jacques (1985). *Noise: The Political Economy of Music*. Minneapolis: University of Minnesota Press.
- Cavarero, Adriana (2005). *For More Than One Voice: Toward a Philosophy of Vocal Expression*. Stanford, Calif: Stanford University Press. Print.

- Dolar, Mladen** (2006). *A Voice and Nothing More*. Cambridge, Mass: MIT Press.
- Downie, J. S. , Tcheng, David K., and Xiang, Xin** (2008). "Novel interface services for bioacoustic digital libraries" in Proc. 8th ACM/IEEE-CS Joint Conf. on Digital Libraries. New York: ACM, 2008: 423-423.
- Fuller, Matthew.** (2005). *Media Ecologies: Materialist Energies in Art and Technoculture*. Cambridge, Mass: MIT Press.
- Gitelman, Lisa.** (2005) *Always Already New: Media, History, and the Data of Culture*. Cambridge, Mass: MIT Press.
- Hirschkind, Charles.** (2006) *The Ethical Soundscape: Cassette Sermons and Islamic Counterpublics*. New York: Columbia University Press.
- High Performance Sound Technologies for Access and Scholarship* project blogs.ischool.utexas.edu/hipstas/
- Hughes, Thomas Parke.** (1983) *Networks of Power: Electrification in Western Society, 1880-1930*. Baltimore: Johns Hopkins University Press. Print.
- Josephson, Matthew** (1992). *Edison: A Biography*. New York: J. Wiley. Print.
- Kirschenbaum, Matthew** (2008). *Mechanisms: New Media and the Forensic Imagination*. Cambridge, Mass: MIT Press.
- LaBelle, Brandon** (2010). *Acoustic Territories: Sound Culture and Everyday Life*. New York: Continuum.
- Mintjes, Louise** (2003). *Sound of Africa!: Making Music Zulu in a South African Studio*. Durham: Duke University Press.
- Morton, David.** (2004). *Sound Recording: The Life Story of a Technology*. Westport, CT: Greenwood Press. Print.
- Moten, Fred** (2003). *In the Break: The Aesthetics of the Black Radical Tradition*. Minneapolis: University of Minnesota Press.
- Oard, D., M. Wu, K. Kraus, A. Hajj-Ahmad, H. Su, R. Garg** (2014). "It's about Time: Projecting Temporal Metadata for Historically Significant Recordings." Forthcoming, Proceedings of the 2014 iConference. Berlin, Germany. 4-7 March 2014. ACM Digital Library. Web.
- Ochoa, Ana Maira** (2006). "Sonic Transculturation, Epistemologies of Purification and the Aural Public Sphere in Latin America." *Social Identities* 12.6: 803-25.
- Punyasena, Surangi W., Tcheng, David K., Wesseln, Cassandra, Mueller, Pietra G** (2012). "Classifying black and white spruce pollen using layered machine learning." *New Phytologist* 196.3: 937-944.
- Rodgers, Tara** (2010). *Pink Noises: Women on Electronic Music and Sound*. Durham, NC: Duke University Press.
- Smith, Mark M** (2006). *How Race Is Made: Slavery, Segregation, and the Senses*. Chapel Hill: University of North Carolina Press. Print
- Smith, Mark M** (2001). *Listening to Nineteenth-Century America*. Chapel Hill: University of North Carolina Press. Print.
- Sterne, Jonathan**. *The Audible Past: Cultural Origins of Sound Reproduction*. Durham, NC: Duke University Press, 2003.
- Sterne, Jonathan.** (2011) "Audio in Digital Humanities Authorship: A Roadmap." (essay in progress) Super bon! Online: superbon.net/?p=1915 . Accessed June 7, 2013.
- Sterne, Jonathan** (2012). *MP3: The Meaning of a Format*. Durham: Duke University Press.
- Su, H., Garg, R., Hajj-Ahmad, A., Min Wu** (2013). "ENF Analysis on Recaptured Audio Recordings." Proceedings of the 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Vancouver, BC, Canada. 26-31 May 2013. 3018-3022. Web.
- Thompson, Emily Ann** (2002). *The Soundscape of Modernity: Architectural Acoustics and the Culture of Listening in America, 1900-1933*. Cambridge, Mass. : MIT Press.
- Toop, David** (2010). *Sinister Resonance: The Mediumship of the Listener*. New York: Continuum.
- Weheliye, Alexander G.** (2005). *Phonographies: Grooves in Sonic Afro-Modernity*. Durham, NC: Duke University Press.
- For cultural histories of aurality, see entries for Smith 2001, 2006 in the bibliography; for sound reproduction, Sterne 2003, 2012; for the politics of the voice, Cavarero 2005, Dolar 2006; and noise, Attali 1984; for urban soundscapes, see entries for Toop 2010, Thompson 2002, Labelle 2010; ethnographic modernities and acoustemologies are covered in Hershkoff

2006 and Ochoa 2006. The sonic construction of gender receives treatment in Rodgers 2010 and Martin 1991; for race and ethnicity, see Weheliye 2002, Moten 2003, Smith 2006, and Meintjes 2003.

New and recent developments in image analysis: theory and practice

Crowther, Charles

Ioannou Centre for Classical and Byzantine Studies, Oxford University, UK

Nyhan, Julianne

Department of Information Studies, UCL London, UK

Tarte, Segolene

Oxford e-Research Centre, UK, United Kingdom

Dahl, Jacob

Oriental Institute, Oxford University, UK

New and recent developments in image analysis: theory and practice

1. Introduction - David Robey, Oxford e-Research Centre

The historical concentration on text in humanities computing and the Digital Humanities (DH) partly reflects the technologies that have been available, and partly the majority interests of humanities researchers. Yet much humanities research also depends on scholars' visual skills, not only in the arts and archaeology, but also in disciplines whose main concern is text, for whom the physical form of texts can be as important as their content. Thus digital textual resources increasingly link to images, thereby greatly increasing their potential scholarly benefits: an outstanding example is Prue Shaw's recent digital edition of Dante's Divine Comedy. This panel session is concerned with what one might consider the next stage in the use of image in DH research: the technologies that are increasingly being used for image recognition, enhancement and analysis. Unlike many of the digital text-and-image archives now available, which do little more than accelerate and facilitate humanities research, these enable the production of knowledge that would simply not be accessible by non-digital means.

The panel presents a range of innovative humanities research in progress in these areas of image recognition, enhancement and analysis at Oxford University and UCL London. It includes both practical applications of the related technologies, and a more theoretical approach to the methods deployed. The latter is an area in which the humanities have traditionally been weak. Scholars have usually been reluctant to reflect in any depth on the exact nature of their methods; even during the heady days of humanities theory towards the end of the last century, the focus of interest was much more on high-level concepts, and much less on the details of methodology. Yet this kind of reflection is essential: if the most productive use is to be made of the new technologies for image analysis, and indeed all other forms of digital analysis in the humanities, and if we are to promote their use effectively, we need to be very clear exactly what they help us to do, and how this fits in with the work that scholars have traditionally done.

The panel will begin with a presentation by Charles Crowther, from the Centre for the Study of Ancient Documents at Oxford, of a range of techniques to increase the legibility of different forms of writing in the ancient world—ancient world studies being probably the field in which scholarship depends most on analysing the exact form in which text is preserved, and as a result the field in which the use of these technologies is most advanced. This will be followed by an in-depth presentation and discussion by Jacob Dahl, of Oxford's Oriental Studies

Institute, of the use of one particular form of these technologies, Reflectance Transformation Imaging, to advance the study of one of the world's earliest and still undeciphered writing systems, proto-Elamite. The paper by Segolene Tarte, from the Oxford e-Research Centre, will provide the theoretical dimension by identifying the cognitive processes involved in some of research covered by the first two papers, and in other related work. Finally Julianne Nyhan, from the Department of Information Studies at UCL London, moves into a more difficult and experimental area with a review of the use of image recognition for historical research and a brief presentation of a project for the digital study of newspaper photographs in the context of the history of the First World War.

The work that the four papers deal with is highly detailed and specialist, but has potential applications far beyond the fields in which it has been carried out so far: a topic we plan to cover in the panel discussion.

* * *

2. Reading Ancient Writing: Technology and Scholarship - Charles Crowther, Centre for the Study of Ancient Documents, Oxford

Much of the evidence that the scholarly community in Classics has available to extend and renew its fields of investigation is fragmentary, difficult to decipher, and tantalising. The use of new technologies opens the prospect of making this evidence more easily and extensively accessible and exemplifies the contribution of DH to scholarly research in a well-defined and coherent context.

In this paper I review the effectiveness of a range of visualisation technologies deployed to increase the legibility of ancient, primarily Greek and Latin, incised and inscribed documentary texts, and consider some directions for future work. The analysis draws perspectives from work in this field at the Oxford Centre for the Study of Ancient Documents (CSAD) over the last 15 years, and presents results from recent and continuing projects undertaken with other presenters in the panel.

I consider two types of text that offer challenges to decipherment that are broadly similar but different in significant respects: wooden and metal writing tablets and inscriptions on stone.

Regular discoveries of incised wooden and metal writing tablets in excavation of Romano-British (and Northern European) sites potentially offer new categories of evidence, but their transcription and decipherment present the constant challenge of separating fine traces of writing from background features and, in many cases, from other palimpsest layers of text. The great majority of the material is relatively new (recovered since 1980) and is still in the early stages of integration into the body of research resources in Ancient History. Successive projects undertaken since 1998 at CSAD, in collaboration with colleagues in the field of medical image analysis in the Department of Engineering Science at Oxford, have resulted in the creation of new techniques for improving the visibility and legibility of writing on wooden tablets, principally by means of a stroke detection method (shadow stereo or phase congruency) and the removal of wood-grain (Brady et al. 2005).

These advances were based on digital scans of the writing tablets made with lights illuminating the surfaces from different angles, calibrated manually. Their application has already resulted in new editions of texts of writing tablets from the Roman fort at Vindolanda (Bowman and Tomlin 2005) and, most strikingly, a Roman legal document found in Frisia in 1914 (Bowman et al. 2009), which forms a case study in the paper offered by Dr. Tarte. Central to this programme of research has been the belief that in order to develop better imaging techniques we need at the same time to explore developments in the representation of semantics, in theories of reading, and in ideas about knowledge representation (Terras 2007; Tarte 2011).

Inscriptions on stone are one of the most characteristic legacies of the culture of the ancient Greco-Roman world,

from the beginning of alphabetic writing in the 8th century BC to Late Antiquity. Very large quantities of inscriptions have been recorded – the total number now published exceeds 800,000 – but few have survived intact; stones are frequently broken into fragments and very many have suffered extensive surface damage from abrasion or erosion. Techniques for the decipherment of these damaged texts had not until recently advanced significantly since the beginning of epigraphic studies in the 15th century. The principal traditional means of reading letter traces, by taking paper (or latex) casts of the surface which can be manipulated more easily than the original stone, or by using solutions of charcoal and water to emphasise surface indentations, remain effective, but involve direct action upon the surface of the stone and are now permitted only under controlled conditions and in exceptional cases by museum conservators.

Because the language and formal character of inscribed documents are well understood, small improvements in reading can lead to significant advances in decipherment and interpretation. Two examples may be cited: much of the history of the 5th-century Athenian empire has seemed to turn on the interpretation of a handful of evanescent letter traces on a stele recording an alliance between Athens and Segesta (Chambers et al. 1990); a palimpsest inscription on a basalt stele recovered during the rescue excavations at Zeugma in the Euphrates valley in 2000 has provided new insights into one of the more remarkable expressions of ruler cult in antiquity (Crowther 2013).

However techniques for recovering text, whether based upon paper casts, illumination with raking light, or, more recently, laser scanning can only improve legibility when there are some remaining topographic traces of the original inscribed text. In this section of the paper I summarise the results of experiments using the microfocus spectroscopy beamline at the Diamond Light Synchrotron in 2010 and 2011, following earlier work at the Cornell High-Energy Synchrotron Source (CHESS), which show that trace elements associated with wear of the inscribing tool and with the pigments used to paint inscribed letters can be detected with high sensitivity and spatial resolution by X-Ray Fluorescence (XRF) Imaging, even when the stone surface has worn below its original contours (Powers et al. 2005).

XRF imaging, for the moment, requires that text artefacts be brought to a synchrotron source and is ineffective where the surface of the object has been subject to intensive cleaning since Antiquity. In the great majority of cases, analysis of surviving surface traces continues to be the principal method of decipherment. Approaches to incised and inscribed texts, accordingly, converge on the need to recover and interpret surface topography as accurately as possible. The manually calibrated methods of illumination used in previous work at CSAD have now been replaced by Reflectance Transformation Imaging (RTI), whose application to cuneiform texts and seal impressions is described in the paper by Dr. Dahl.

RTI uses multiple images captured from a fixed camera position to construct a digital model of surface form and reflectance for the object studied. The resulting files enable interactive changes to lighting, image enhancements and automated identification of visual and morphological attributes. RTI has a number of specific advantages for the capture of incised and inscribed documents: non-contact acquisition of surface data, to alleviate the concerns of museum conservators; potential representation of 3D shape characteristics without data loss due to shadows and specular highlights; virtualised surface analysis under any form and distribution of lighting; the possibility of analysing surfaces remotely and 're-photographing' them for dissemination. RTI representations of documentary texts are visually striking and attractive, but current fitting algorithms do not exploit the full potential of the image data captured. In the final section of the paper I report on current work undertaken at CSAD to improve the algorithms and capture processes and their application to a range of incised and inscribed documentary texts.

Selected References

- A. K. Bowman, R. S. O. Tomlin**, "Wooden Stilus Tablets from Roman Britain", in Bowman and Brady 2005, 7-14
- A. K. Bowman, M. Brady** (eds.) *Images and Artefacts of the Ancient World* (British Academy Occasional Papers, Oxford 2005)
- A. K. Bowman, R. S. O. Tomlin, K. A. Worp**, "Emptio bovis Frisica': the 'Frisian ox sale' reconsidered", *JRS* 99, 2009, 156-70
- M. Brady, X.-B. Pan, V. Schenk, M. Terras, P. Robertson, N. Molton**, "Shadow Stereo, Image Filtering, and Constraint Propagation", in Bowman and Brady 2005, 15-30
- M. Chambers, R. Gallucci, P. Spanos**, "Athens' Alliance with Egesta in the Year of Antiphon", *ZPE* 83, 1990, 38-63
- C. Crowther**, "Inscriptions on Stone", in W. Aylward (ed.), *Excavation at Zeugma*. Conducted by Oxford Archaeology (Los Altos, 2013), 192-219
- J. Powers, N. Dimitrova, R. Huang, D.-M. Smilgies, D. H. Bilderback, K. Clinton, and R. E. Thorne**, "X-ray fluorescence recovers writing from ancient inscriptions", *Zeitschrift für Papyrologie und Epigraphik* 152, 2005, 221-227
- S. M. Tarte**, "Papyrological Investigations: Transferring Perception and Interpretation into the Digital World", *Lit Linguist Computing* 26(2), 2011, 233-247.
- M. Terras**, *Image to Interpretation. An Intelligent System to Aid Historians in Reading the Vindolanda Texts* (Oxford 2007)

Reflectance Transformation Imaging of ancient texts – Jacob Dahl Oriental Studies Institute, Oxford

This paper explores the use of Reflectance Transformation Imaging (RTI) technology in the study and decipherment of ancient texts. RTI is "a computational photographic method that captures a subject's surface shape and color and enables the interactive re-lighting of the subject from any direction" (definition from <http://culturalheritageimaging.org/Technologies/RTI/>). It is based on the work of Tom Malzbender (2001, see <http://www.hpl.hp.com/research/ptm/index.html>).

The camera dome used at the University of Oxford, and built by researchers at the University of Southampton (Earl et al. 2011), uses 76 daylight-LEDs, which are attached to the inside of the plexiglas dome. A high-resolution digital camera is mounted on the top of the dome, looking straight down through a hole. The object is placed on a stage in the centre lifted up to the horizon. The diameter of the camera dome is approximately 1m, allowing for the capture of objects up to 33 cm in diameter. 76 individual raw files are captured each using a different light source and therefore a different light angle. In post-processing the images are joined to create a composite image (model) (Polynomial Texture Mapping (PTM)) where the light-source can be changed by the user.

Using RTI images captured in the Louvre Museum in Paris, researchers at the university of Oxford have been able to significantly advance the study of one of the world's earliest and still undeciphered writing systems, proto-Elamite, mimicking in the classroom the work of the epigrapher in the museum. This method has proven particularly valuable when examining secondary additions to certain signs, lightly impressed signs, alterations to signs, or seal impressions.

Proto-Elamite is the conventional name given to a derived writing system emerging in Iran following the spread of the culture and technological advances of the Late Uruk period in Mesopotamia into Western Iran c. 3500 BC (Dahl 2013a). The writing system is defined by having a high number of singletons, and possibly a high number of scribal errors, perhaps resulting from never having been standardized (Dahl 2002 and Dahl 2013b). A majority of the extant proto-Elamite texts are kept in the Louvre Museum, Paris, and the National Museum of Iran, Teheran. The writing system disappeared after a short use of at most a few centuries, and writing was not used in Iran for the following five centuries or longer.

It has long been realized that deciphering early scripts involves more than merely a linguistic puzzle, that features such as seal impressions, scribal marks, etc. hold valuable information, and that the materiality of writing is therefore more

important for the study of early writing than anywhere else (Damerow 2006). Subtle differences in sign forms may be the result of scribal hands, semantic variation, or simply lack of practice (André-Salvini and Dahl in press).

Traditional print-representations of early writing specimens only poorly represent the physicality of the object. Proto-Elamite and other early writing systems are often studied by very small groups of researchers at universities or research institutions across the globe. Previous generations of researchers were confined to either using hand copies of the originals, of varying quality, or consulting the originals in museums far from their home institutions, when attempting to decode the information of these documents. High-resolution, dynamic images of these text artifacts therefore have the potential of transforming the study of early writing by simulating first-hand consultation of the originals, enabling shared research, and bringing together disparate data-sets.

Over the course of the last two years c. 1100 tablets and fragments in the Louvre Museum were imaged with the camera dome (André-Salvini and Dahl 2013a). Results of research facilitated by these images is now being published. For example, the study of RTI images of two tablets in the Louvre Museum (Sb 15229 and 15456) challenged the existing view of the imagery of the seal impression found on both (a couple of humanoid figures as well as animals), and led to a strengthening of the theory that no representations of the human form was allowed in glyptic art of the proto-Elamite period (results later confirmed by collation of originals in the Museum) (Dahl 2014).

The main issue facing a wider application of RTI technology in the study of ancient writing is the size of the files (mostly 256 mb per image, six needed for a cuneiform tablet), and the lack of a suitable on-line viewer. Image size is becoming less of an issue over time, and the issue of an on-line viewer is the focus of at least one ongoing project (<http://www.arts.kuleuven.be/info/ONO/Meso/cuneiformcollection> and <http://portablelightdome.wordpress.com>). In the meantime captures taken by the camera dome can be used to produce very high quality static images by blending different views together in an image editor.

Selected References

- André-Salvini and Dahl** 2013: "L'écriture proto-élamite: la numérisation à l'aide du déchiffrement des premières écritures". In *Grande gallerie: le journal du Louvre* 2013/2, 28-29.
- André-Salvini and Dahl** in press: "Le système d'écriture proto-élamite : un point sur le déchiffrement des signes, leur formation et leur usage". In *La recherche au musée du Louvre* 2012 (2013).
- Dahl** 2002: "Proto-Elamite Sign Frequencies." In *CDLB* 2002:1 (2002) http://cdli.ucla.edu/pubs/cdlb/2002/cdlb2002_001.html
- Dahl** 2013a: "Early Writing in Iran". In Potts, D.T., *Oxford Handbook of Iranian Archaeology* (Oxford University Press) (2013), 233-262.
- Dahl** 2013b: "Frühe Schrift im Iran". In N. Crüsemann et al. *Uruk 5000 Jahre Megacity: Begleitband zur Ausstellung "Uruk–5000 Jahre Megacity"* im Pergamonmuseum–Staatliche Museen zu Berlin (2013), 202-203.
- Dahl** 2014: "The proto-Elamite seal MDP 16 no 198". In *CDLN* 2014:1 <<http://cdli.ucla.edu/Pubs/cdln/archives/000028.html>>.
- Damerow** 2006: "The Origins of Writing as a Problem of Historical Epistemology." In *CDLJ* 2006:1 (2006) <http://cdli.ucla.edu/pubs/cdlj/2006/cdlj2006_001.html>
- * * *

Digital Images of Ancient Textual Artefacts. Connecting Computational Processing and Cognitive Processes - Ségolène Tarte, Oxford e-Research Centre

Digital image processing is an expanding domain in the DH that naturally finds its place in the overall knowledge and meaning creation process that is the ultimate aim of the study of ancient textual artefacts. The cognitive aspects of this intrinsically interpretative process play a major role in the endeavour and continually interact with the computational processing mobilized by image capture and processing methodologies. In this paper, I aim to present some of the cognitive aspects of the interpretation of textual artefacts that intervene in the analysis of their digital image avatars. To illustrate those cognitive processes, I will present them in the context of "naturalistic" observations of expert papyrologists and assyriologists working on ancient textual artefacts (observations made following ethnographic methodologies), and connect them with similar observations made in the controlled settings of laboratory experiments as reported in the cognitive sciences literature. Each example illustrates how one specific cognitive process has been aided by the use of digital image-based technology, and how they are integrated into the interpretation workflow.

It is worth specifying that the processes highlighted here are perceptual in nature, rather than conceptual; however, as they participate in the act of interpretation of textual artefacts, in the act of knowledge creation and sense-making, they definitely qualify as cognitive processes. These cognitive processes are efficient and complementary with the services that digital tools can render. My aim in identifying them is not to attempt to emulate them digitally, but rather to identify where computational processing can provide help and where the upper hand is best left to the experts.

Making the intangible tangible: Artemidorus papyrus (in collaboration with Prof. D'Alessio (KCL), and Dr Elsner (Oxford)). Here the computational processing involved infrared image capture, and, in later work, digital image alignment (of the front and back images) as well as virtual rolling of the papyrus (Tarte, 2012). The virtual rolling was made in order to evaluate the hypothesis that the papyrus fragments needed to be reordered based on traces of inks from the reverse of the papyrus that seemed to have transferred to the obverse by mirror impression while the papyrus was rolled up. Beyond providing a rigorous argument in favour of reordering, the actual process of virtually rolling the papyrus prompted a re-materialization of it. A physical avatar was produced which allowed the researchers to experience for themselves the rolling of the document and thereby ascertain that the ink transfers could have resulted from the roll, confirming the plausibility of the reordering of the fragments. This enacted approach to interpretation, enabled by the upstream image-based technologies that have been mobilized, points to what the neurosciences call **embodied cognition**, where "Social Meaning is primarily the object of practical concern and not of theoretical judgment... It relies on non-inferential mechanisms, which do not require the explicit use of rationality" (Gallese, 2005, p43). Through a combination of image processing output and enacted engagement with a physical avatar, aspects of the intangible papyrus have been made tangible.

Making the inarticulate articulate: the Roman stylus wooden tablet known as the "Frisian Ox" tablet (in collaboration with Prof. Bowman (Oxford) and Prof. Terras (UCL)). Beyond image capture, here the computational processing involved removing the woodgrain and enhancing the visibility of the scratches that constitute the script (Tarte, 2011). Digital technologies were also used to produce line drawings by means of a drawing tablet that allowed for the tracing of the text over any digital image from the collection that had been captured. One of the cognitive processes that ensued has to do with **visual completion**. By tracing the letter shapes, experts filled in the gaps where portions of character were absent. This enacted approach to interpretation, enabled by the upstream image-based technologies that have been mobilized, points to the phenomenon the neurosciences call **illusory contour**: "Detection of an illusory figure shows a precedence of specific global object properties over local attributes ... it is the surface rather than the contour that guides search" (Conci et al., 2007, p1293-4). This in part explains why the expert papyrologists stipulated that the woodgrain removal algorithms applied to the images were not only not very helpful, but also possibly

confusing. Through a combination of illusory contour detection and digital tracing of the letters, aspects of the inarticulate text of the tablet have been made articulate.

Making the invisible visible: proto-Elamite tablets from ancient Iran (in collaboration with Dr Dahl (Oxford) – cf paper on this panel). Here the computational processing involved the deployment of an advanced image capture technique known as Reflectance Transformation Imaging (RTI) (Earl et al., 2011)). Through this technique, it is possible to interactively change the position of a unidirectional light source shone onto the artefact, thereby allowing for the accentuation of its physical geography - an enormous asset when dealing with 3D scripts such as proto-Elamite. In effect, what RTI allows is to mobilize **depth perception through monocular motion parallax** (Rogers and Graham, 1982): RTI supports depth perception, and results in making visible the otherwise invisible.

Making the indiscernible discernible: Selenite curse tablets from Ancient Cyprus (in collaboration with Dr Bodard (KCL) and Prof. Radaelli (Oxford)). Here the crystalline nature of the support makes the texts indiscernible. As mounting evidence in recent bodies of work in the cognitive sciences points to an **action-simulation-perception framework** (related to embodied cognition), where visualizing the results of an action mobilizes the pre-motor areas of the brain that correspond to the observed action (such as in reading/writing, or in painting (Taylor et al., 2012)), Scanning Electron Microscope (SEM) images, which reveal disruptions in the crystalline structure of selenite, have the potential to inform the viewer on the ductus and the dynamics of the act of writing, thereby facilitating reading by making the indiscernible dynamics of writing discernible.

In each of these examples I have selected one specific cognitive process that took place. Of course, many others occurred, and each of those cases could have been used to illustrate one or more of the other cognitive processes highlighted here. Further work will explore the interaction of the perceptual and conceptual cognitive processes mobilized in the interpretation of textual artefacts, and how to integrate them in a digitally supported workflow.

Selected References

- Conci, M., Müller, H. J., and Elliott, M. A. (2007). *The contrasting impact of global and local object attributes on kanizsa figure detection*. Perception & Psychophysics, 69(8):1278—1294.
- D'Alessio, G. (2009). *On the "Artemidorus" Papyrus*. Zeitschrift für Papyrologie und Epigraphik, 171:27—43.
- Earl, G., Basford, P. J., Bischoff, A. S., Bowman, A., Crowther, C., Hodgson, M., Martinez, K., Isaksen, L., Pagi, H., Piquette, K. E., and Kotoula, E. (2011). *Reflectance transformation imaging systems for ancient documentary artefacts*. In Electronic Visualisation and the Arts, London, UK. BCS, The Chartered Institute for IT.
- Gallese, V. (2005). *Embodied simulation: From neurons to phenomenal experience*. Phenomenology and the Cognitive Sciences, 4:23—48. 10.1007/s11097-005-4737-z.
- Rogers, B. and Graham, M. (1982). *Similarities between motion parallax and stereopsis in human depth perception*. Vision Research, 22(2):261 — 270.
- Tarte, S. M. (2011). *Papyrological Investigations: Transferring Perception and Interpretation into the Digital World*. Lit Linguist Computing, 26(2):233—47.
- Tarte, S. M. (2012). *The digital existence of words and pictures: The case of the Artemidorus papyrus*. Historia, 61(3):325—336 (+bibliog. pp 357—61; fig. pp 363—5).
- Taylor, J. E. T., Witt, J. K., and Grimaldi, P. J. (2012). *Uncovering the connection between artist and audience: Viewing painted brushstrokes evokes corresponding action representations in the observer*. Cognition, 125(1):26 — 36.

Facial recognition and Digital Humanities: new directions? - Julianne Nyhan, Department of Information Studies, UCL

As documentary sources photographs can offer historians new ways to uncover, interrogate, visualise and communicate about the past. In the past new insights into historical questions have been gained through unexpected and often serendipitous observations in historical photographs: for example, the seemingly accidental discovery of the young Hitler in photographs of social-democratic rallies (cf Krumreich 2001). Recent developments in facial recognition and image analysis techniques (see, inter alia, Wang et al. 2013, Singh et al. 2013 and Vieira et al. 2013) offer historians (and researchers across the humanities and beyond) new ways to think about and analyse visual documentary evidence. The application of facial recognition techniques may well contribute to the establishment of a more systematic basis for the discovery and analysis of actors and social networks in historical photographs, and perhaps for automatically identifying 'suspect' photographs. Nevertheless, comparatively little research seems to be ongoing within the DH into the applications of such facial detection, recognition and visual computing techniques.

Automated facial detection has become ubiquitous in day to day life. It is used on public and private CCTV-networks for crime and terror prevention, for example, by preselecting information for screening and at border controls using biometric passports. Digital cameras and mobile phones can auto-detect faces and even smiles on faces (see Deniz, O. et al. 2008), replacing traditional passwords or providing 'convisual' information to photographers, like the names of the people in the viewfinder if these people have been photographed before (see, for example, Brown 2011). Social media networks like Facebook or Twitter engage in the batch-tagging of people in photographs, while popular image cataloguing software packages like Google's Picasa allow automatic graphical indexation of large snapshot collections at home. Automated facial detection, despite still being a relatively new branch of image analysis, has quickly matured and allows the facial indexation of large graphical databases. The result is that graphical information can be searched in ways that have hitherto been impossible.

Nevertheless, within the context of DH, few applications of such techniques can be noticed in the published literature. The paper will start with a review of literature related to image analysis techniques in order to present both the technological state of the art and uses that are being made of image analysis techniques in DH. A review of the literature relating to historians' methodological engagement with visual documentary sources will also identify existing and new research problems in historical research that facial recognition techniques could be applied to, leading to a discussion of the potentials and drawbacks that facial recognition techniques hold for this kind of research.

In the context of DH, the most substantial research to date is that of Suárez, de la Rosa Pérez and Ulloa (2013), who have applied such techniques to representations of the human face in world literature. To do so they analysed more than 123,500 paintings from all periods of art history with a face recognition algorithm used in Facebook's photo-tagging system and automatically identified over 26,000 faces. Using techniques like, among others, the Elastic Bunch Graph Matching (Wiskott et al. 1997), they were able to detect what they term basic features (e.g. information about the position of facial features such as eyes and nose) and extended features (e.g. gender, mood, age range). They conclude that by comparing "the basic features set using graphs and the extended features set using clustering by K-Means method (Sculley 2010) ... we are at the perfect position to analyze and characterize each of the groups according to different historical perspectives and cultural questions, for instance, the distinction among styles by giving a minimum set of features that determines its membership" (p, 535).

From the perspective of facial recognition techniques noteworthy recent developments include the 3D face recognition systems by the University of Bradford's Centre for Visual Computing, which they state has led to the development of a prototype system that "demonstrates that 3D facial data can overcome many inherent problems in image-based (2D) face recognition. This can be accurate up to the level of differentiating between identical twins" (see University of

Bradford 2013). However, for my present purpose it is mostly historical photographs that are under discussion and at the time of writing it is not clear how such a system would work with historical photographs (even if 3D models of the 2D photographs were created using technology such as freely available software platforms like 123D Catch).

To conclude the talk an overview of the initial findings of a project on First World War photographs that we have recently started will be presented. WW1 was the first major war where photography was affordable and routinely used in both official and amateur channels; thus visual documentary evidence from the period is vital to its study. During WW1 Belgian refugees arrived in Britain en masse in what transpired to be the largest ingress of refugees in British history. However, our understanding of who arrived, how they intersected with British and diasporic social networks, how long they stayed, and whether they settled or returned home is limited. One of the historical applications reported on in this paper will be the initial findings of a pilot project that is investigating accuracy rates of facial detection techniques on historical photographs of WW1 Belgian refugees.

References

- Brown, Mark.** 2011. "Facebook Silently Rolls Out Face Recognition Tagging to the World." *Wired*, 08 June 2011 edition, sec. Technology. <http://www.wired.co.uk/news/archive/2011-06/08/facebook-face-recognition>.
- Deniz, O., Castrillon, M., Lorenzo, J., Anton, L., and G. Bueno.** (2008) 'Smile Detection for User Interfaces', Advances in Visual Computing, vol. 5359/2008, pp. 602-611
- Krumreich, Gerd.** "Hitler in der Menge," in: C. Dipper, A. Gestrich, L.Raphael (eds.), *Krieg, Frieden und Demokratie: Festschrift für Martin Vogt zum 65. Geburtstag* (Frankfurt, 2001), 137-140.
- Singh, Chandan, Ekta Walia, and Neerja Mittal.** 2012. "Robust Two-stage Face Recognition Approach Using Global and Local Features." *The Visual Computer* 28 (11) (November 1): 1085–1098. doi:10.1007/s00371-011-0659-7.
- Suárez, de la Rosa Pérez and Ulloa** (2013) 'Not Exactly Prima Facie: Understanding the Representation of the Human Through the Analysis of Faces in World Painting' DH Book of Abstracts. 534-6. Nebraska: Centre for Research in the Humanities
- University of Bradford,** '3D Face Recognition For Security Systems ' www.visual-computing.brad.ac.uk/case-studies/3d-face-recognition-security-systems (accessed 31/10/2013)
- Vieira, Tiago F., Andrea Bottino, Aldo Laurentini, and Matteo De Simone.** 2013. "Detecting Siblings in Image Pairs." *The Visual Computer*: 1–13. doi:10.1007/s00371-013-0884-3.
- Wang, Zhifei, Zhenjiang Miao, Q. M. Jonathan Wu, Yanli Wan, and Zhen Tang.** 2013. "Low-resolution Face Recognition: a Review." *The Visual Computer*: 1–28. doi:10.1007/s00371-013-0861-x.

Remediating 20th-Century Magazines of the Arts: Approaches, Methods, Possibilities

Ermolaev, Natalia

Princeton University, United States of Americ

Wulfman, Clifford E.

Princeton University, United States of Americ

Biber, Hanno

Austrian Academy of Sciences

Crombez, Thomas

Royal Academy of Fine Arts Antwerp

Remediating 20th-Century Magazines of the Arts: Approaches, Methods, Possibilities

1. Panel Proposal

The remediation of historical books and newspapers into the digital environment receives ample attention in today's academic and library communities, and has become a core feature of major digital humanities projects that have enabled innovative methods of inquiry and new scholarly discoveries. However, despite the establishment of pioneering digital resources such as the Modernist Journals Project and the Modernist Magazines Project, the design and research potential of electronic collections of modernist historical magazines remains an understudied and under-theorized topic.

This panel aims to fill this gap by presenting a series of approaches, methodologies and possibilities inherent in the creation of robust digital collections of 20-th century magazines of the arts. We bring together designers of three collections – the Blue Mountain Project (Princeton University), the AAC-Austrian Academy Corpus (Austrian Academy of Sciences), and Digital Archive of Belgian Neo-Avant-garde Periodicals (or DABNAP, at the Royal Academy of Fine Arts Antwerp) – for a conversation addressing both the conceptual underpinnings as well as the practical applications of their work. This international panel will continue a lively discussion started at a conference at Princeton University in October 2013 on remediating avant-garde magazines. Our goal at DH 2014 will be to illustrate a variety of avenues available for digital curation of historical magazine collections, and to move toward a set of shared guidelines for representing this rich material in the electronic environment and facilitating advanced research.

A discussion of the theoretical and practical issues involved in remediating modernist and avant- garde periodicals is both important and timely. In recent years, as the field of Periodical Studies has cohered as subset of print culture scholarship, researchers have started looking at historical periodicals in new ways and discovering that they fundamentally challenge our conventional understandings of modern culture. The recent publication of the Oxford Critical and Cultural History of Modernist Magazines, a hefty three-volume set of essays that discusses over 500 magazines from Europe and the Americas, underscores the relevance of periodicals for today's research on modern and modernist culture. If, as Sean Latham and Robert Scholes assert in "The Rise of Periodical Studies," the flourishing of this field has been enabled and invigorated by the affordances of the digital environment – what is the response of the Digital Humanities community to the continued evolution of Periodical Studies? What are our obligations as designers of collections of digital historical periodicals? How must we build resources that embody, allow, and promote the sophisticated new avenues of scholarly engagement made possible by electronic tools and platforms?

The panelists' answers to these questions will reflect the scope, methodology, and focus of their projects: Crombez's DABNAP is concerned with magazine as a performance and art text, Biber's paper illustrates the corpus linguistics approach, and Wulfman and Ermolaev maintain a Periodical Studies perspective. Each will touch upon the intellectual and technological insights that emerge when we remediate the arts magazine, such as: representation of aesthetic, material, and social features, questions regarding materiality (format, typography, paper, binding, etc.), tracing printing and distribution history, data modeling, linguistic analysis, semantic enrichment and tagging, entity and name recognition, interface design, and application and tool-building.

This panel promises to stimulate a coherent discussion that will be informative to a broad range of digital humanists. By presenting a cross-section of types of collections, the panel will demonstrate the current state of innovative projects on 20-th century arts magazines, and help forge the way for future work.

All panelists have agreed to participate.

2. Paper 1: Magazines of Magazines. Corpus Research Applications for New Digital Editions of Historical Magazines (Hanno Biber)

A magazine can be regarded as a specific container of texts. A corpus can be regarded as a structured and complex collection of texts. The simple questions of how to structure these complex texts in text corpora and how to integrate magazines in corpora and to make them accessible for research purposes is particularly challenging. This paper will present corpus research applications for innovative digital editions of magazines. Therefore, three research aspects have to be considered. First, the methodological parameters of corpus research have to be determined and the applicability of corpus linguistics for the question of designing and building digital text editions of historical magazines has to be examined. Second, the research potential of scholarly digital editions of magazines in the context of a digital humanities framework has to be described. And third, a practical corpus-based methodological approach of magazine studies has to be given by investigating the specific textual qualities of the magazines and the specific language use in the magazines' texts, which have been annotated and made accessible with the help of corpus linguistics as well as innovative interface design and graphic design principles. The study will present several aspects of these research questions illustrated by text examples and discuss the methodological implications of such corpus-based investigations into the use of language in texts in particular. Literary magazines in particular have specific properties that can be recognized and registered by means of a corpus-based study of language. Therefore the literary magazines subsection of the "AAC-Austrian Academy Corpus" will be used as examples for this type of research. This resource is an ideal basis for a corpus linguistic exploration into the study of literary magazines. Corpus-based text studies can be regarded also as instruments of textual critique, whereby the corpus-based approach allows various ways of philological research and text analysis.

The framework of the "AAC-Austrian Academy Corpus" offers a research platform for corpus linguistics in a very broad sense as well as for corpus-based magazine studies in particular. So far, two model digital editions of magazines have been developed within the AAC for the purpose of analysing large amounts of literary texts with a methodological approach offered by corpus linguistics. The "AAC-FACKEL"^[6] and the "Brenner online"^[7] editions are two examples of how corpus research initiatives can help to develop new applications in the field of digital humanities, in particular for language studies in the context of literature research as well as for historical studies and for broader issues related to cultural studies. The "AAC-Austrian Academy Corpus" has been established as a corpus research initiative concerned with exploring electronic text corpora and conducting research in the fields of corpus linguistics, text analysis and digital text corpora. More than 500 million running words of text have been scanned, converted into machine-readable text and carefully annotated with structural mark-up. The texts that have been integrated into the AAC are German language texts of historical and cultural significance. The historical period covered is ranging from 1848 to 1989, a period showing historical changes with remarkable influences on the language and the language use. In the context of the subset of the AAC digital editions the language of satire written by Karl Kraus and his use of language in his magazine is of particular interest for a study on language use and in particular for a study on language use based upon the principles of corpus linguistics. Apart from the digital editions of the "Brenner online" and the "AAC-FACKEL", which has been annotated to a large extent, the AAC magazines corpus holdings provide a great number of reliable resources and interesting corpus based approaches for investigations into the properties of these texts. Several other magazines have been digitized making use of XML-related standards. Both the „AAC-FACKEL“ and „Brenner online“ offer fully searchable online editions of the journal with various indexes and search tools in a web interface, where all pages of the original are available as digital texts and as facsimile images. The „AAC-FACKEL“ and „Brenner online“

allow new methods of scholarly research and philological analysis of texts that are of crucial importance for the history and study of not only the language but also of the status and the properties of the magazine. The edition interface of the AAC digital editions has several sophisticated search mechanisms and indexes as well as five individual frames synchronized within one single window. The philological principles of scholarly digital editions within the "AAC-Austrian Academy Corpus" are determined by the conviction that the methods of corpus research enable valuable text resources and research tools for linguists and literary scholars. The resource is an interesting text basis for corpus linguistic explorations whereby a corpus-based approach allows new ways of philological research and analysis of magazines.

3. Paper 2: The Blue Mountain Project and the Language of Avant-Garde Magazines (Wulfman and Ermolaev)

Introduction

This paper initiates a dialog with Hanno Biber's contribution to this panel by suggesting that a magazine is not only a container of texts but a text itself. As Carolyn Ulrich, one of the authors of the influential book, *The Little Magazine: A History and a Bibliography*, wrote to her co-author Frederick Hoffman, "a magazine is a tricky individual": tricky in its identity and tricky in its individuality. What is a magazine? Is it a particular issue (a manifestation at a particular point in time – a copy, an issuance, etc.), or is it a title with some sort of identity over time (expressed through editorial consistency, name continuity, etc.)?

To ask these questions is to engage Biber's research approach to the magazine – specifically, "the applicability of corpus linguistics [to] the question of designing and building digital text editions of historical magazines." While the methods of corpus linguistics are, without a doubt, of crucial importance to the study of historical magazines, other disciplinary methodologies can compliment lexical analysis by highlighting additional key aspects of this diverse and rich material.

In their 2010 book, *Modernism in the Magazines: An Introduction*, Wulfman and Scholes critique the conventional distinctions that have, historically, been applied to magazines, especially generic labels such as "little", "mass", "literary", "avant-garde", etc. They suggest employing a more analytic approach instead, one which entails identifying a set of characteristics that contribute to our understanding of magazines and then clustering them in different ways, much as linguists characterize speech sounds by clustering features into phonemes. The Blue Mountain Project at Princeton University is continuing this line of inquiry, and in this paper we describe our development of a language of magazines, and of a technological infrastructure and set of applications to represent this complex language in a robust digital resource.

An Ontology of Historical Magazines

The Blue Mountain Project, based in the Princeton University Library and funded by the National Endowment for the Humanities, was launched in 2012 as a freely-available electronic resource for art, music, and literary periodicals published in Europe and the United States between 1850-1923. In the first 2-year grant cycle, Blue Mountain is making 34 titles (approximately 95,000 pages) available in French, German, English, Italian, Spanish, Czech, Russian, Polish, Finnish, and Danish.

In the first part of this paper, we present our research on an ontology of historical magazines – that is, a set of concepts with which knowledge of magazines is represented, expressed in a vocabulary that denotes the types, properties and interrelationships of those concepts – that has become the core intellectual infrastructure of the Blue Mountain Project. This ontology differs from the terms of descriptive bibliography developed in the library community such as MARC (MACHINE-

Readable Cataloging), and from the critical terminology traditionally used by scholars and critics. The purpose of our ontology is to provide a framework on which researchers and scholars may encode and describe historical magazines.

To be sure, some of the concepts do come from the language of descriptive bibliography: title, editor, contributor, and so forth; but others are more nuanced than descriptive bibliography usually expresses. Issuance, for example, is often a complex phenomenon, especially for magazines published internationally, making the concept of an original copy from which a digital edition is derived a vexed one and requiring scholars to rethink narratives of publication. Other concepts, like circulation and price, are tied to data that are difficult to obtain; still others, like readership, are complex concepts whose sub-concepts must be teased apart and identified before they may be used meaningfully.

Other terms in the Blue Mountain ontology come from the languages of typography, book history, and graphic design, and are inspired by Jerome McGann's notion of the "bibliographic code." In order to discuss the relationship of content elements in a magazine, for example – such as the relationship of advertisements to articles – one must have a common language for expressing layout. Simple text transcription of a magazine's contents is insufficient for many kinds of research; thus our ontology is based on an understanding of the historical language of page composition (columns, paragraphs, various forms of headings, publication metadata, and so on) that is vital to the useful encoding of magazine structure and the analysis of a magazine's meaning.

Advanced Applications: Blue Mountaineer and Blue Mountain Springs

In the second part of our paper, we present an experimental architecture for encoding and expressing the language of magazines. The Blue Mountain Project has been designed to support a variety of research uses, beyond the now-standard modes of full-text searching and page browsing. To support those uses, we create high-resolution digital images and provide robust library-standard metadata including title, issue and constituent-level MODS (Metadata Object Description Schema) and METS/ALTO (Metadata Encoding and Transmission Standard/Analyzed Layout and Text Object) records. In our next phase of work, however, we plan to expose this highly structured data for mining and analysis by means of two new modules: Blue Mountaineer and Blue Mountain Springs.

Blue Mountaineer

The Blue Mountaineer is a set of web applications we will design for exploring Blue Mountain content through visualizations, topic modeling, and other forms of data mining and retrieval. It is intended to showcase the power and utility of a rich XML-encoded database while providing researchers and students with tools they can use in their own investigations. Examples of the type of research Blue Mountaineer will enable include the following:

Social Network Analysis: Users will be able to explore the complex publication webs with tools that enable them to plot graphs of relations among titles, authors, artists, languages, and nationalities using Blue Mountain's rich metadata. They will, for example, be able to see all the issues in which Tristan Tzara and Francis Picabia appeared together.

Data-Driven Timelines and Clusters: Blue Mountain will employ off-the-shelf natural-language processing software such as Apache OpenNLP and the Stanford Named Entity Recognizer, as well as Princeton's own WordNet, to perform first-order named-entity detection on its corpus of magazines. The results of these analyses will be encoded in Blue Mountain's TEI transcriptions, which will make it possible for researchers to discover and visualize sequences and relationships buried deep in the textual data itself. They will be able to compare how two authors use a particular term, for

example, or see how the work of a particular artist relates to particular advertisers.

Topic Modeling: The aggregation and clustering of content zones encoded in METS/ALTO make it possible for topic modeling algorithms to perform much more fine-grained analysis of magazines and newspapers than they can perform using unzoned, "dirty-OCR'ed" texts. Researchers will be able to study and compare the abstract topics occurring in the work of two authors, for example, or to compare the topics in a magazine's articles with those in its advertising.

Blue Mountain Springs

Information scientists and digital humanities researchers often want to bypass reader-oriented interfaces and access full-text data directly and programmatically for use with external analytic tools. The Blue Mountain Springs module will make Blue Mountain an abundant source of clean data by providing an application programming interface (API) to Blue Mountain's metadata and machine-readable full-text transcriptions. Blue Mountain Springs will support traditional metadata harvesting via OAI-PMH[14], as well as the more elaborate aggregations supported by OAI-ORE[15]. It will enable software clients to access the plain text of Blue Mountain's materials through web-addressable text streams that can be piped directly into visualization and analysis applications.

4. Paper 3: The Document as Event. Analyzing Artist Networks through a Digital Archive of Avant-garde Periodicals (Crombez)

Introduction

In this panel presentation, I would like to highlight the ideas, problems and outcomes of DABNAP, which stands for the 'Digital Archive of Belgian Neo-Avant-garde Periodicals.' A considerable collection of forty artist periodicals from the 1950s to the 1980s is being digitized, in order to examine the underlying network of artists and artist groups.

My main interest will be the issue of semantic enrichment of the source documents. To various degrees, semantic enrichment is already common to many scholarly digitization projects. Think, for instance, of marking up personal names and locations in a TEI-encoded document. But for large-scale projects, manual semantic enrichment is often unfeasible. Can automatic procedures, such as named entity recognition (NER), help to mine art periodicals? Furthermore, can we imagine and develop software that detects not merely names, but also artistic information in vast collections of text?

In order to answer these ambitious questions, I will first develop a new conceptual model (centered on 'the document as event'), and then highlight the actual context in which this model will be put in practice.

The Document as Event

The archive of back issues from *LIFE Magazine* (spanning five decades) was a showcase project for Google Books when the initiative started in 2004. Apart from access through browsing or through the ubiquitous search box, the internet company uses basic text analysis to help users navigate the archive. The interface presents a cloud of words and expressions that are characteristic for the issue that the user is currently browsing. For the February 1937 issue, the web archive is happy to inform the user that "Leon Trotsky," "Reichstag," "Studebaker," and "Kleenex" are among the common terms and phrases. But it is obviously blind to the question whether these names belong to people, organizations, places, or brands.

Current digital archives, then, create at least as many problems as they solve. Users prefer to direct their questions to a simple search box. But the limitations of this model for search are obvious to academic users, and have recently led Google

itself to introduce the 'Knowledge Graph', enriching search results with semantic metadata. The Knowledge Graph or, more generically, *semantic* enrichment, shows the future direction of digital text collections.

In order to deal with this new conceptual reality behind current and future digitization projects, let me introduce a new conceptual model to think about such document collections.

I would like to conceptualize the *document as event*, in order to transfer something of the dynamics of the event onto the document, which is commonly conceived of rather statically. A periodical is occasioned by artistic events, such as the publication of new literary works, the exhibition of visual art, or the presentation of a new theatrical performance. However, as a document, it can also be considered to be an 'event' in itself. The text functions as a linguistic meeting space for a wide diversity of named entities. This includes names of artists, writers, dramatists, performers, directors and critics; names of museums, galleries, theatres, companies and schools; and titles of books, art works and theatrical productions. In other words, the concept of the document as event serves as a bridge between the literary and art-historical approach to sources as autonomous documents (and hence all too easily viewed as unconnected), and the linguistic approach to sources as text (which all too easily flattens the document).

The DABNAP Collection

The artistic renewal in Belgium since the 1950s, sustained by small groups of artists, led to a first generation of postwar artist periodicals. Titles such as *Le surréalisme révolutionnaire*, *Cobra*, *De Tafelronde*, *Het Cahier* and *Gard Sivik* proved decisive for the formation of the Belgian neo-avant-garde in literature and the visual arts.

During the 1960s and the 1970s, happening and socially engaged art (inspired particularly by the Provo movement) took over and gave a new orientation to artist periodicals. Examples include *Happening News*, *Revo*, *Anar*, *Milky Way*, *Total's*, and, on the side of literature, *Labris*, *Yang*, *Bok*, *MEP*, *Heibel*, *Boemerang*, and many others. Finally, the 1970s and 1980s saw the rise of punk-inspired zines, including *Force Mental*.

The challenges and difficulties of this project lie in dealing with non-standard formats, types of paper, typography, and non-paper inserts. Paper sizes range from the ludicrously large (A2) to the very small (half of A5). Printing techniques include offset, mimeograph, screen printing and photocopy – resulting in extremely diverse kinds of lettering and typography, which often confuses the OCR software that is used to extract text from the scanned pages.

Apart from the technical difficulties of scanning and extracting text from the heterogeneous source documents, further difficulties of the DABNAP project include interface design and handling copyright issues, which will briefly be discussed in the presentation. The main and final difficulty (or rather, ambition) of DABNAP is to process of extracting complex information about artistic events from the text. This will require, first, to expand common procedures for named entity recognition with techniques for recognizing titles and events. In other words, which means are appropriate, and which linguistic tools have to be developed, for the task of recognizing meaningful relationships between names (e.g., that a certain director is the author of a theatre production)?

References

Both large-scale digitization of books (such as Google Books and the Open Content Alliance) as well as small, curated collections (such as ArchBook <http://archbook.itschool.utoronto.ca>) are transforming reading and research practices in literary studies and book history. **Ryan Cordell**'s research using data from the Library of Congress "Chronicling America" project is a prime example of new scholarly engagement with historical newspapers (see Cordell's "Uncovering Reprinting Networks in Nineteenth-Century American Newspapers," <http://www.viralteats.org>).

Modernist Journals Project: <http://modjourn.org>; *Modernist Magazines Project:* <http://www.modernistmagazines.com>.

The 19th century periodical has been more heavily studied; see especially the project, nineteenth-century serials edition (<http://www.ncse.ac.uk>) and the essays: **James Mussell and Suzanne Paylor**, "Mapping the 'Mighty Maze': The Nineteenth-Century Serials Edition," *Nineteen: Interdisciplinary Studies in Nineteenth-Century Studies*, 1 (2005) and **Mussell, Paylor, M. Deegan and K. Sutherland**, "Editions and Archives: Textual Editing and the Nineteenth-Century Serials Edition (ncse)," in *Text Editing, Print, and the Digital World* (Farnham: Ashgate, 2009), 137-158.

See **Sean Latham and Robert Scholes**, "The Rise of Periodical Studies," *PMLA*, Vol. 121, No. 2 (Mar., 2006), pp. 517-531. In this article, Latham and Scholes discuss the context of North American academia.

The Oxford Critical and Cultural History of Modernist Magazines: Volume I: Britain and Ireland 1880-1955 (ed. Peter Brooker and Andrew Thacker, Oxford University Press, 2009); *Volume II: North America 1894-1960* (ed. Brooker and Thacker, 2012); *Volume III: Europe 1880 - 1940* (ed. Peter Brooker, Sascha Bru, Andrew Thacker, Christian Weikop, 2013).

AAC-Austrian Academy Corpus: AAC-FACKEL. Online Version: »*Die Fackel*. Herausgeber: Karl Kraus, Wien 1899-1936«. AAC Digital Edition No 1, (Editors-in-chief: Hanno Biber, Evelyn Breiteneder, Heinrich Kabas, Karlheinz Mörtl; Graphic Design: Anne Burdick) <http://www.aac.ac.at/fackel>.

AAC-Austrian Academy Corpus und Brenner-Archiv: BRENNER ONLINE. Online Version: »Der Brenner. Herausgeber: Ludwig Ficker, Innsbruck 1910-1954«. AAC Digital Edition No 2, (Editors-in-chief: Hanno Biber, Evelyn Breiteneder, Heinrich Kabas, Karlheinz Mörtl; Graphic Design: Anne Burdick) <http://www.aac.ac.at/brenner>.

Robert Scholes and Clifford Wulfman, *Modernism in the Magazines: An Introduction* (New Haven, CT: Yale University Press, 2010)

Modernist Journals Project: <http://modjourn.org>; *Modernist Magazines Project:* <http://www.modernistmagazines.com>.

Jerome McGann, *The Textual Condition* (Princeton, NJ: Princeton University Press, 1991)

<http://opennlp.apache.org>
stanford.edu/software/CRF-NER.shtml
<http://wordnet.princeton.edu>
<http://www.openarchives.org/pmh>
<http://www.openarchives.org/ore>

Global Outlook::Digital Humanities: Promoting Digital Humanities Research Across disciplines, regions, and cultures

O'Donnell, Daniel Paul
 University of Lethbridge, Canada

Bordalejo, Barbara
 University of Saskatchewan

Risam, Roopika
 Salem State University

Spence, Paul
 King's College, London

González-Blanco, Elena
 UNED, Spain

Global Outlook::Digital Humanities: Promoting
Digital Humanities Research Across disciplines,
regions, and cultures

Daniel Paul O'Donnell University of Lethbridge

By DH 2014, Global Outlook::Digital Humanities (GO::DH) will be 18 months old. Old enough to have experienced its first growing pains but also old enough to have a sense of the opportunities that exist for promoting the globalisation of Digital Humanities research. This paper discusses the past and future of the Special Interest Group (SIG), concentrating particularly on what is generalisable: the lessons we have learned about collaborating in a multilingual, multiregional context (although this makes it in some sense a project report, the lessons are themselves highly relevant to and have implications for the community as a whole).

The basic premise behind GO::DH is both exciting and frightening. It is a community for Digital Humanities scholars around the world that encourages them to discover new work and new colleagues in regions and disciplines they might never have considered before. But in asking scholars to do this, it also asks them to work along some of our most controversial lines of division: language, culture, nationality, history, and income level. In a field that is famously and self-consciously "nice" (Scheinfeldt 2010), these are the places where our self-conception has been tested (and called into question) most vigorously (see especially Fiormonte 2012 on language and nationality; Babalola 2012 discusses some of the challenges that divide the use of the digital in High Income vs. Low Income Economies).

The English language and (arguably at least) Anglo-American disciplinary and rhetorical norms dominate the practice of our profession. This places non-native speakers of English and scholars working outside the Anglo-American academic context at an immediate disadvantage. In addition, as Fiormonte suggests, it probably also has led to the relative scarcity of such scholars in the disciplines'gatekeeping positions.

The interest in technology that defines our field, moreover, creates divisions the moment our collaborations attempt to cross the boundaries that distinguish high, mid, and low income economies (O'Donnell 2012a). As Babalola has shown, the kind of basic infrastructure that Digital Humanities scholars in High Income Economies take for granted either does not exist or can be disproportionately difficult and expensive to access in mid- and especially low-income economies. As she demonstrates, moreover, this problem is about more than download speeds or CPUs: many of our core approaches, assumptions, and methods of dissemination (from the outsized importance of conference presentations in our discipline to the use of crowdsourcing) imply an access to resources common only in High Income economies.

And finally there is the spectre of colonialism and development politics. Any project that brings scholars from high-income economies into close contact with scholars from mid and low income economies is going to run into questions of intention, history, and politics. Can such collaborations be collaborations of equals, in which all participants both teach and learn? Or must they inevitably resolve themselves into the more unequal relationship of donor and recipient? The initial impetus for GO::DH arose among scholars working in High Income economies who wondered about their lack of contact with scholars working in other regions (O'Donnell 2012b). Initially, this caused suspicion among scholars who live in or work with those in mid and low-income regions. What was the motivation for interest from the high-income scholars? To what extent would this new organisation be able to avoid replicating the status quo in the field at large, where those with resources determine the course of the collective effort.

Despite our initial fears about what could go wrong in such an endeavour, the first year of GO::DH's existence has been remarkably productive and relatively smooth. While there have been some misunderstandings (some of which are discussed in the other papers in the panel), there have been remarkably few problems. The SIG has successfully managed to integrate multi-lingualism into its discussion-list, which several threads being carried out in languages other than English. It has provided a framework for a remarkable number of projects and working groups—from Around the World of DH to the second Caribbean ThaTCamp, to the first Global DH conference (planned for this coming Spring in Mexico in association with

RedHD. And it has even led to the formation of new groups, such as the proposed Portuguese-language DH organisation.

The techniques we have used in building this community, capturing the good will of its constituents while avoiding some of the more obvious potential problems offers wider lessons for the DH community. In this talk I will discuss some of the specific techniques—from ad hoc translation to the collaborative development of by-laws and executive positions that we have used to successfully build GO::DH over the last year into the relatively stable community it has now become.

Works cited

- Babalola, Titilola.** (2012). *"The Digital Humanities and Digital Literacy: A Review of Digital Culture in Nigeria".* Lethbridge, Alta.
- Fiormonte, Domenico.** (2012). *"Towards a Cultural Critique of the Digital Humanities."* Historische Sozialforschung / Historical Social Research 37 (3) (September): 59–76.
- O'Donnell, Daniel Paul.** (2012a). *"In a Rich Man's World: Global DH?"* Dpod Blog. November 2. dpod.kakelbont.ca/2012/11/02/in-a-rich-mans-world-global-dh .
- O'Donnell, Daniel Paul.** (2012b). *"Global Outlook :: Digital Humanities"*. Lethbridge: Alliance of Digital Humanities Organisations. ubuntuone.com/187LiVZpJKwFNaRV0IZJeD .
- Scheinfeldt, Tom.** (2010). *"Why Digital Humanities Is 'Nice'."* Found History. May 26. www.foundhistory.org/2010/05/26/why-digital-humanities-is-%e2%80%9cnice%e2%80%9d .

Thinking with an Accent

Barbara Bordalejo, University of Saskatchewan

This paper explores the reasons behind various degrees of impact and development in Digital Humanities in the context of different countries, languages and cultures. It calls attention to the enormous gap between low, middle and high-income countries and offers avenues to continue a change that has already begun. Firstly, the paper considers where is Digital Humanities located as a discipline and how this might influence the phenomenon of inclusion/exclusion (particularly those that refer to language). Secondly, it focuses on practical problems in countries of low and middle-income. Thirdly, it suggests practical ways to construct a field that is truly inclusive and allows wider participation.

Locating the Digital Humanities

Matthew Kirschenbaum, in his article “What is Digital Humanities and What is Doing in English Departments?,” puts forward several possible reasons that explain why English Departments are one of the main spaces in which Digital Humanities is cultured. As Kirschenbaum explains it, the phenomenon is related to text as a relatively easy object to encode (“...by far the most tractable data type for computers to manipulate.”), the relationship between computers and composition, the relationship to editorial theory and the work by Jerome McGann in the 1980s, the advent of electronic literature, “the openness of English Departments to cultural studies,” and the development of e-readers, which have finally made it possible to have digital texts in the same form as we have digital music.

It seems that English and History are the main areas in which Digital Humanities posts are advertised. However, Roopika Risam (roopikarisam.com/2013/09/15/where-have-all-the-dh-jobs-gone) has pointed out that most of the recent DH jobs “go hand-in-hand with Rhetoric and Composition and literature positions.” In practical terms, this means that there is some substance to the Kirschenbaum claim about Digital Humanities being easily located within English Departments (generally responsible for the teaching of composition and solid in the teaching of literature). If we, temporarily, accept this premise as true (that the English Department is a place in which we can find an important cluster of Digital Humanists) we have to

consider the consequences of this: English Departments might have the natural tendency of hiring English Native speakers. An examination of the structure of the Alliance for Digital Humanities Organizations offers an insight on the distribution of power in Digital Humanities: in the steering committee, of 27 positions only Elisabeth Burr, Christoph Meister, Masashiro Shimoda, Oyvind Eide and Edward Vanhoufte are not British, North American (or more generally, native English speakers, as is Paul Arthur, from the Australasian Association for Digital Humanities). This is an observation of fact and not meant as a criticism of ADHO. In paper and pixels, the policy is of inclusion. ADHO has a committee in multilingualism and multiculturalism. However, those who come from the fringes either know or suspect that equity cannot be reached by decree (if this were possible, we would only have to pass laws forbidding poverty and the issue would be solved).

The Myth of Openness in DH

The state of affairs is surprising because it is widely believed among the high ranking digital humanists 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 (59). Notably, this statement by Kirschenbaum is also supported by Burdick et al. (“...however heterogeneous, the Digital Humanities is unified by its emphasis on making, connecting, interpreting, and collaborating” (24)), among others. 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 make us unable to see that the discipline fails to meet these standards.

Global Outlook :: Digital Humanities

Through GO::DH we have been exploring these and other issues and it is becoming clear that the discipline has a clear center and well defined margins. In April 2013, Frédéric Clavert published a note entitled “The Digital Humanities multicultural revolution that did not happen yet” (www.clavert.net/the-digital-humanities-multicultural-revolution-did-not-happen-yet). Domenico Fiormonte published the link and started a thread in which scholars from very different backgrounds weighed in on their own positions (listserv.uleth.ca/pipermail/globaloutlookdh-l/2013-April/000308.html). Clavert considers the disparity between the number of Anglo-American reviewers for the DH conference and the clear interest in the Francophone community. What defines the field as Anglophone, according to Clavert, is that the “...Digital Humanities, though claiming to be new and revolutionary, are structured in a very classical way for an academic field, where those who master the English language and the English speaking and impact factor based academic journals are the most visible (and the most quoted).” His statements were received with equal amounts of disbelief and approval. As if to confirm Clavert’s position, Craig Bellamy tried to dismiss this issue (listserv.uleth.ca/pipermail/globaloutlookdh-l/2013-April/000309.html). Although the exchange was academic and polite, it brought to light various issues we must face in a global context:

1. There are projects and initiatives being developed of which we are not aware because they are buried in a non-English context.
2. There are cultural factors that affect communication, of which one of the most important ones is the perceived imperialism of the imposition of English as lingua franca.
3. This same linguistic profiling is instrumental in the process of exclusion to which non-native speakers are subject because of the lack of native abilities.

At least, Fiormonte will agree with me in saying that a specific agenda to exclude non-native speakers from Digital Humanities is unnecessary: our linguistic limitations already prevent us from being considered central. Fiormonte’s example of Dino Buzzetti’s rise as an authority in the field (something that occurred not because of his excellent work, or the thirty

years of publications in English and Italian, but rather because of the support of recognized scholars) painfully shows the lack of openness of this supposedly all-embracing field. As Fiornonte says: "...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 (listserv.uleth.ca/pipermail/globaloutlookdh-l/2013-May/000329.html).

Of course, the question here is how can this problem be solved. It is clear that we cannot both blame the establishment and ask for a solution coming from it. As I stated before, this is not a problem of ill will, but rather a misunderstanding of the determining factors that create these problems.

Minding the Gap

As a member of the executive of GO::DH, I consider that my role is to identify our problems to generate solutions to them. By organizing Spanish language THATCamps and delivering DH lectures in Argentina and Mexico, I have come to understand that to level the field we need not only to translate texts that currently are only available in English (the Cátedra Datos at the Universidad de Buenos Aires does exactly that: a team of seven people produce translations of up to date texts required as an introduction to Digital Humanities), but we also need to understand the reluctance of scholars to subject themselves to English as the exclusive language for communication and we have to allow for non-native standards to be considered when submissions to conferences or journals are made by non-native speakers. We should not forget that those who speak or think with a foreign accent, are able, at least, to speak another language.

The Possibilities and Pitfalls of Global Digital Humanities

Roopika Risam, Salem State University

Over the past two years, much has been made of the role of cultural critique in the digital humanities, particularly around silences and absences of race, gender, sexuality, and so forth in the digital humanities (Liu 2012; McPherson 2012; #transformDH Collective 2011; Lothian and Phillips 2013; Bailey 2011). Yet, these conversations have taken shape through a United States-centric frame of reference that often elides the larger picture of the digital humanities: its global frame. Taking up the global scope of the digital humanities, however, is to take up imbalances of power that operate in colonialist frames, visible in the dominance of United States and Western European voices within the digital humanities community writ large. Indeed, it requires heightened attention to cultural critique through a postcolonialist framework.

Reactions within the digital humanities community to cultural critique (Whitson 2012, Reid 2011) renders such critiques as problems. Indeed, the problem, as the narrative goes, lies not in gaps within the digital humanities but with the practitioners who dare to raise these issues. In this talk, I will examine the "problem" of the global in the digital humanities. I begin by outlining the stakes of attending to global participation in the digital humanities. These stakes are both intangible and tangible and include radically reimagining loci of the digital humanities beyond the current map that locates the United States and Western Europe at the center, strategizing models for global partnerships outside of neocolonial frames, and developing resources for fostering a truly global digital humanities.

As the stakes imply, attending to the global within the digital humanities requires a two-pronged approach that accounts for the complexities of both theory and praxis. Engaging these concerns, my talk provides a case study of theoretical and practical approaches: Global Outlook :: Digital Humanities (GO::DH) and Postcolonial Digital Humanities (DHPoco). GO::DH is a special interest group (SIG) of the Alliance of Digital Humanities Organizations. GO::DH is dedicated to hacking barriers that prohibit collaboration across

both disciplines and geographies (Global Outlook :: Digital Humanities 2013). Barriers include telecommunications, financial resources, human labor, and language (O'Donnell). Focusing on stated goals of "discovery, community-building, research, and advocacy," GO::DH works to foster communication and collaboration on a global scale. DHPoco is both a movement and emergent subfield of the digital humanities, invested in decolonizing digital spaces, making space for colonial critique and anti-colonial thought in the digital humanities, and writing alternative genealogies of the digital humanities (DHPoco 2013). By examining the work of GO::DH and DHPoco, I make the case for continued attention to and development of theoretical and organizational spaces for fostering the global digital humanities, as well as its benefits to the digital humanities community as a whole.

References

- Moya Bailey** (2011), "*All Digital Humanists Are White, All Nerds are Men, but Some of Us Are Brave*," *Journal of Digital Humanities* 1.1.
- Alan Liu** (2012), "*Where Is Cultural Criticism in the Digital Humanities?*" Debates in the Digital Humanities. Minneapolis: University of Minnesota Press.
- Tara McPherson** (2012), "*Why Are the Digital Humanities So White?*" Debates in the Digital Humanities. Minneapolis: University of Minnesota Press.
- Daniel O'Donnell** (2012), "*In a Rich Man's World: Global DH?*" 2 November 2012, dpod.kakelbont.ca/2012/11/02/in-a-rich-mans-world-global-dh
- Global Outlook** (2013) : Digital Humanities, "*About*," www.globaloutlookdh.org/about
- Alexis Lothian and Amanda Phillips** (2013), "*Can the Digital Humanities Mean Transformative Critique?*" *Journal of E-Media Studies* 3.1.
- Alex Reid**, "*Alan Liu, Cultural Criticism, the Digital Humanities, and Problem Solving?*"
- Roopika Risam and Adeline Koh** (2013), "*Mission Statement*," Postcolonial Digital Humanities, dhpoco.org/mission-statement-postcolonial-digital-humanities/#transformDH
- Collective, "A Call to Action," 26 October 2011 www.hastac.org/blogs/amanda-phillips/2011/10/26/transformdh-call-action-following-asa-2011
- Roger Whitson, "Does DH Really Need to Be Transformed?," 8 January 2012, www.rogerwhitson.net/?p=1358

Global Challenges, Local Interpretations. An analytical perspective about DH in Spain

Paul Spence (King's College London) and Elena Gonzalez-Blanco (UNED, Spain)

Digital Humanities has a long history in the Spanish-speaking world, with landmark projects like Admyte and BOOST/Philobiblon emerging in the 1970s and 1980s (and later the Miguel de Cervantes Digital Library), followed by years of isolated research projects (in Spain, the focus for this paper, these have often had a strong philological focus, but also encompass bibliographic studies, multimedia and other forms of digital scholarship) and a number of experiences in teaching, including the now defunct online Masters programme in Digital Humanities at the University of Castilla-La Mancha, Spain, which ran with some success for a few years. But as is common with non-Anglophone traditions, the rich history in digital humanities in Spain is under-represented internationally, and in this particular case has not even achieved a consistent institutional presence in Spain. 2013 was a milestone in the history of digital humanities in the Spanish language (Baraibar 2013), and Spain saw a number of events and initiatives,iv including the inaugural conference ('Digital Humanities: challenges, achievements and future perspectives') in July 2013 of the newly-formed association *Humanidades Digitales Hispánicas* (HDH, 'Hispanic Digital Humanities in English'). HDH joins a broader articulation of

Hispanophone digital humanities organisations, which started with the Mexican association *RedHD*, and reflects a broader flowering of initiatives in Spain, which however do not conform a homogenous whole.

A survey of digital humanities in Spain

The term ‘humanidades digitales’ is a trending topic in Spain, and in this paper we explore its manifestations, its tensions and its challenges. A wide-ranging history of digital humanities in Spain is much needed, both to communicate the historic and disciplinary depth of the field in the country and to help give substance to the development of the field in Spain itself. What is needed as a prelude to this, however, is stable documentation of the field as it currently stands, and what we have done is to carry out a broad survey of digital humanities activities in Spain with a view to making it available for further research by others. In our research, we have surveyed conferences, publications, official and unofficial websites and blogs in Spain, in addition to broader international resources.

Digital humanities in Spain still suffers from relative invisibility at an international level in digital humanities, but the evidence changes in different settings – for example, Spanish representation is relatively prominent in the results of the ‘Who are you Digital Humanists?’ survey carried out by researchers from OpenEdition, with 40 respondents (compared to 49 in the UK, which has a population 30% higher). In part this is due to more general issues with how digital humanities is defined and represented internationally, but self-identification is a major issue – many researchers are interested in the digital humanities without necessarily feeling themselves to be represented by the label. While there is a relatively strong theoretical tradition often connected to conventional humanities disciplines (digital philology, digital art) or information science, there is not a strong history of tool-building that is more prevalent in other regional contexts.

Communities, definitions, labels

Domain-specific communities have played important role in developing awareness of digital scholarship, although they may not self-identify as digital humanities entities – some of the greatest progress has been made in groups such as TC/12, a research project with major funding to explore texts and research tools in Spanish Golden Age theatre, or CHARTA, an international research network which both provides guidelines for editing Spanish archival texts from the twelfth to nineteenth centuries, although in some cases the engagement with technology is uneven or still under negotiation (Spence et al 2012). Sometimes digital humanities finds its expression more comfortably in MediaLabs, as is the case of the MediaLab of Salamanca, which recently organised a series of seminars, or broader scholarly initiatives in social and human sciences, such as GRINUgr, which examines digital humanities from the prism of digital culture.

If in 2006 Isabelle Leibrandt could ask if Digital Humanities was a science fiction or an imminent reality (Leibrandt 2006), no-one could argue of its existence as such now. The question, and not only in a Spanish context of course, is what is it? Is it just a convenient label which allows each person to project their own fantasies, as Olivier le Deuff puts it (Le Deuff 2012), or is it a set of fully-formed academic practices? Labels may not be particular illuminating here, but the pattern in Spain, where any label has been used at all, is to use the term ‘informática humanística’, roughly equivalent to ‘humanities computing’ in English, and which is probably most closely influenced by the Italian usage ‘informatica umanistica’. Unlike in Italian, where the historic term has persisted (as evidenced in the name of the Italian association, evidence of the use of the term largely disappears in around 2008, when we gradually see the emergence of ‘humanidades digitales’, an almost direct translation of ‘digital humanities’.

In a blog on the relationship of the digital humanities to information science, Luis Rodriguez-Yunta asks why we use the

term ‘digital humanities’ and not just ‘digital scholarship’. In his own response, he points to the academic, social and cultural demand for accessible and humanities-focused sources/documentation, but also, crucially, the role of the humanities in defence of the human, implying a ‘humanisation’ of technology.

Spain has suffered especially badly during the global economic crisis, with exceptionally high unemployment figures and drastic cut in funding, which has in turn heightened the sense of crisis in the academy, where the humanities are perceived as being especially vulnerable to criticism, and some have perceived this crisis as an opportunity to redraw traditional disciplinary lines. The Digital humanities have a strong background in the philological tradition in Spain – indeed the two initial seminars which led to the creation of the HDH association in Deusto and La Coruña with strong philological characters. But there are also strong voices for a recalibration of the humanities, which in the words of José Manuel Lucía, can use digital technologies to recuperate a social space it gradually lost in the twentieth century (Lucía, 2012).

Focus for the future

The HDH conference in July 2013 was the focus for a number of key debates affecting Spanish digital humanities at this time, including the role of teaching and systems of academic value and credit. The HDH association has filled an important void in Spain, providing formal structures for digital humanities, offering an organisational focus and functioning as a mechanism to lobby national academic institutions responsible for the formal evaluation process. We will end our paper by exploring the crucial role of HDH and other initiatives in helping the digital humanities to establish itself in Spain, and describe efforts to create an academic centre of digital humanities in Spain, which up until now has not existed, with a focus on research, teaching and general support for digital humanities practitioners.

References

- Baraibar, Álvaro** (2013) ‘Buenos tiempos para las Humanidades Digitales en español’ Blog dhd2013.filos.unam.mx/porvistadeojos/2013/05/20/buenos-tiempos-para-las-humanidades-digitales-en-espanol
- Dacos, Marin** (2013) ‘La stratégie du Sauna finlandais’ Blog blog.homo-numericus.net/article11138.html
- Galina Russell, Isabel** (2011) ‘¿Qué son las Humanidades Digitales?’ in Revista Digital Universitaria Vol. 12, No.7, www.revista.unam.mx/vol.12/num7/art68/index.html
- González-Blanco, Elena** (forthcoming). ‘Las Humanidades Digitales vistas desde España’
- Le Deuff, Olivier.** (2012) *Humanisme numérique et littératies*, Semen n° 34, p.117-134
- Leibrandt Isabel** “Humanidades, ciencia ficción o realidad inminente?” www.ucm.es/info/espacio/numero33/humadigi.html
- Lucía, José Manuel** (2012). *Elogio del texto*, Madrid, Fórcola
- Romero, Esteban** ‘Humanidades Digitales en investigación, docencia y universidad’ presentation at estebanromero.com/2013/10/humanidades-digitales-en-investigacion-docencia-y-universidad
- Spence, P., Isasi Martínez, C., Pierazzo, E. & Vincente Miguel, I.** (2012) *Nuevas perspectivas para la edición y el estudio de documentos hispánicos antiguos*. Sánchez-Prieto Borja, P. & Torrens Alvarez, M. J. (eds.). Bern: Peter Lang
- Spence, Paul** (forthcoming). Report on first Digital Humanities conference in Spain (forthcoming in Japanese)
- Zotero group for ‘humanidades digitales’ https://www.zotero.org/groups/humanidades_digitales
- www.admyte.com
- bancroft.berkeley.edu/philobiblon/history_en.html iii
- www.cervantesvirtual.com
- hd.paulspence.org/recursos/hh-dd-es/
- hdd2013.humanidadesdigitales.org
- www.humanidadesdigitales.com
- humanidadesdigitales.net

tc12.uv.es
www.charta.es
medialab.usal.es
grinugr.org
Associazione per l'informatica umanistica e la cultura digitale,
AIUCD www.umanisticadigitale.it

Spectacle vivant et technologie numérique: du laboratoire scientifique au plateau de théâtre

Pluta, Izabella

Izabella.E.Pluta@gmail.com

Ecole Nationale Supérieure des Arts et Techniques du Théâtre

Fourmentraux, Jean-Paul

Université de Lille 3

Bardiot, Clarisse

Université de Valenciennes

Spectacle vivant et technologie numérique: du laboratoire scientifique au plateau de théâtre

Artistes de l'ère numérique. Art, science ou technologie ?

de Jean-Paul Fourmentraux

Qu'est-ce que « créer » dans un contexte interdisciplinaire hybride arts, sciences et technologies numériques ? Depuis une dizaine d'années le numérique bouscule les frontières entre des domaines de l'activité artistique qui étaient jusque-là relativement cloisonnés : arts plastiques, littérature, spectacle vivant, musique et audiovisuel. Nombre de projets artistiques en lien avec les technologies informatiques et multimédias mettent en œuvre des partenariats pluridisciplinaires où cohabitent le théâtre, la danse, le cinéma ou la vidéo et le son.

La création artistique et la recherche technologique, qui constituaient autrefois des domaines nettement séparés et quasiment imperméables, sont aujourd'hui à ce point intriqués que toute innovation au sein de l'un intéresse (et infléchit) le développement de l'autre. Les œuvres hybrides qui résultent de leur interpénétration rendent irréversible le morcellement des anciennes frontières opposant art et science. La manière inédite dont celles-ci se recomposent amène à s'interroger d'une part sur l'articulation qui, désormais, permet à la recherche et à la création d'interagir, et d'autre part sur la redéfinition des figures de l'artiste ainsi que des modes de valorisation des œuvres spécifiques à ce contexte. Car plus que de transformer seulement les modalités du travail de création, un enjeu tout aussi important de ces partenariats réside dans la nécessaire redéfinition de la (ou des) finalité(s) de ce qui y est produit. La question cruciale devenant alors celle de la clôture de l'œuvre et de ses mises en valeurs entre logiques artistiques (qualité esthétique, projet d'exposition) et technologiques (recherche et développement, transfert industriel). Le suivi d'*« affaires »* de recherche-création en art numérique (Fourmentraux 2013) révèle en effet des enjeux renouvelés – mutations du travail artistique, redéfinition des modes de production et de circulation des œuvres, outils et stratégies renouvelés de mise en public, en exposition ou en marché – qui entraînent une transformation des modes d'attribution et de valorisation des œuvres, partagées entre art, science et développement technologique.

Création interdisciplinaire et œuvres frontières

L'histoire des interfaces entre art et science s'est pourtant fondée sur un conflit culturel entre des acteurs dont les qualifications, savoirs et compétences, étaient conçues a priori comme opposées. Dans ce contexte, les efforts de recherche et de création mettent en évidence une même contradiction : d'un côté, l'injonction au progrès, encouragé par ces nouvelles industries culturelles numériques qui favorisent le travail en réseau et prêchent pour la reconnaissance de la « valeur créativité » comme nouvel enjeu économique mondial ; d'un autre côté, le durcissement paradoxal des hiérarchies, voire des conflits, entre mondes de l'art et entreprises technologiques. Nombre de partenariats sont encore dirigés vers la production d'une œuvre d'art, ou d'un outil technologique, compris comme des finalités opposées, en conformité avec un régime de propriété exclusive. La division du travail subordonnait le plus souvent l'expérimentation artistique au projet industriel, en privilégiant le développement de nouveaux services ou usages de la technologie. Mais de nouvelles institutions entre art et science voient le jour pour mieux accompagner la transformation des pratiques de recherche et de création : Art Science Factory (Paris-Saclay), Programme doctoral SACRe (PSL et Ensad Paris), Ircam (Paris), Iméra (Marseille), Pictanovo, Imaginarium et Fresnoy (Tourcoing), CEA Minatec et Scène nationale de Meylan (Grenoble), Artem (Nancy), Hexagram (Montréal). L'activité artistique en ressort quelque peu transformée, partagée entre des acteurs multiples qui investissent individuellement et collectivement une œuvre-frontière tendue entre des logiques simultanément artistiques (qualité esthétique, enjeu d'exposition) et technologiques (enjeu de recherche et développement, transfert industriel). Les cas de double réussite restent bien sûr encore rares, mais il en existe : qualité des productions artistiques et de leur rayonnement dans le milieu des arts, doublé d'une mise en marché efficace et rentable d'applications ou de procédés technologiques directement issus de la recherche artistique ou de la production d'œuvres culturelles. Dans ce contexte, la production de « valeurs croisées » ne présuppose pas une synergie de l'art et de l'entreprise. Au contraire, évitant les écueils de la fusion ou de l'instrumentalisation, il s'agit d'organiser la relation dans le sens d'un apprentissage réciproque et d'une production multicentrique.

L'examen de ces croisements de l'innovation artistique et technologique met désormais en jeu une conception coordonnée, un développement agrégé et une valorisation fragmentaire de la production :

- Le travail de conception doit y être coordonné dans la mesure où il met en relation les savoirs et savoir-faire hybrides de collectifs hétérogènes : artistes, ingénieurs, chercheurs.
- La phase de développement doit agréger ces traductions de buts et d'intérêts en un programme de création homogène visant à garantir l'irréversibilité des résultats.
- Mais la valorisation suppose in fine de fragmenter ces résultats pour les redistribuer entre les collectifs et les mondes hétérogènes dans lesquels ils pourront circuler.

Autrement dit, chacun des partenaires - détenteurs de savoirs et de compétences hétérogènes, inscrits dans une culture ou un corps professionnel qui a ses propres valeurs, mais aussi ses instances de désignations et de légitimation spécifiques de ce qu'est le travail, l'œuvre, l'action – y est invité à renouveler le cadre et les modalités de la relation et de l'échange artistique.

En résumé, la recherche-création introduit donc deux critères désormais essentiels :

- le travail en équipes interdisciplinaires ;
- l'impératif d'un programme de recherche transversal à plusieurs œuvres ou projets artistiques.

L'exigence de « valeurs croisées »

Il s'agit en effet de favoriser une certaine « modularité » de la production, en même temps que des formes alternatives de distribution des activités de création et de leurs résultats.

Trois types de projets phares peuvent être distingués :

- les « créations artistiques », qui mènent vers la réalisation d'une œuvre, d'un dispositif ou d'une installation artistique;
- les « découvertes technologiques », qui impliquent le développement de logiciels ou d'outils novateurs;
- les « contributions théoriques », qui poursuivent une perspective analytique et critique d'accumulation de connaissances.

Ce morcellement du travail créatif engendre donc des modes pluriels de désignation de ce qui fait « l'œuvre commune ». Dans ce contexte, la création ne repose plus sur un schéma hiérarchique qui ferait intervenir une distribution réglée des apports en conception et en sous-traitance, selon des échelles de valeur et de rétribution enrôlant une longue chaîne de travailleurs, au service, à chaque fois, d'un créateur singulier. Le travail de création se voit au contraire distribué sur différentes scènes et entre plusieurs acteurs pour lesquels il est possible de préciser des enjeux de recherche distinctifs, suivant des expertises et des agendas variés. L'enjeu vise ainsi un dépassement du « conflit culturel » caractéristique des modèles antérieurs de convergence « arts - sciences - technologies » entre des acteurs (scientifiques, artistes, industriels, amateurs) dont les qualifications, compétences et finalités étaient a priori conçues comme opposées. Gageons que cet élargissement des issues de la recherche artistique et de la création scientifique permettra une plus grande diversité culturelle. En ce sens que les usagers, mais aussi les amateurs d'arts et de science, y gagneront une meilleure compréhension, à la fois sensible et intelligible, des technologies et de leurs enjeux. À mesure que l'activité de création et d'invention se fait de moins en moins monopolistique : loin de présenter une perte, l'interdisciplinarité permettant d'œuvrer non pas à la fusion mais à la confrontation des idées et des émotions, vers un enrichissement réciproque de l'art et de la science.

References

- Dautrey, J.** (2010) (éd.), *La recherche en art(s)*, Paris, MF.
- Entre arts et sciences*, Culture et Musées, n°19, sous la direction de Bordeaux, M.-C., Actes Sud, 2012.
- Fourmentraux J-P.** (2010), *Art et Internet*, Paris, CNRS éditions.
- Fourmentraux J-P.** (2011), *Artistes de laboratoire*, Paris, Hermann.
- Fourmentraux J-P.** (2012), (dir.) *Art et Science*, Paris, CNRS éditions - Les Essentiels d'Hermès.
- Lévy-Leblond J-M.** (2010), *La science n'est pas l'art*, Paris, Hermann.
- Risset J-C.** (1998), *Art, Science, Technologie*, Paris, Rapport de mission MENRT.
- Menger P-M.** (1989), *Les laboratoires de la création musicale*. Paris, La Documentation française.

Le Geminoïde F ou les limites du laboratoire

de Izabella Pluta

Nous souhaitons analyser dans cette communication le parcours technologique et théâtral du Geminoïde F, robot humanoïde conçu par le roboticien Hiroshi Ishiguro dans son Laboratoire ATR à Osaka et mis en jeu théâtral par le metteur en scène et auteur dramatique Oriza Hirata.

La question du robot s'impose de plus en plus dans le spectacle contemporain qui intègre le dispositif technologique. Les solutions robotiques y apparaissent sous de multiples facettes et des réponses esthétiques : une marionnette électronique, un exosquelette, un bras artificiel ou encore un robot anthropomorphe. Ce dernier peut être alors intégré d'une manière différente et faire partie de la scénographie, du dispositif, ou exercer une fonction performative, en se rapprochant de plus en plus du comédien et de son jeu. Il peut accompagner l'interprète sur le plateau, comme nous le verrons dans le travail scénique de Hirata ou encore être le seul acteur dans l'espace désigné comme l'aire de jeu. Ce dernier cas de figure est exploré par une activité artistique nommée le *théâtre robotique*, exercé par Chico MacMurtrie et son collectif

Amorphic Robot Works ou encore par un tout jeune Clément-Marie Mathieu et sa plateforme Thé-Ro, par exemple.

Le Geminoïde F, à son tour, semble aspirer à un véritable rêve de Pygmalion, car il se rapproche de l'apparence humaine comme aucun autre robot conçu jusqu'à présent ne l'a fait. Il représente une belle jeune femme de 25 ans dont le visage trouve son original en une personne réellement existante dont l'identité est confidentielle. Cet androïde nous trompe par sa peau imitant parfaitement la peau humaine ainsi que par sa mimique dotée du mouvement des lèvres et du clignement des paupières. En effet, le Geminoïde F représente une complexité électronique extrêmement avancée, qui est, pour l'instant, le point culminant des recherches menées par le professeur Ishiguro avec son équipe.

Le Geminoïde F surprend également par son parcours original s'étendant entre le laboratoire scientifique et le plateau de théâtre, parcours qui semble s'élargir encore vers les domaines de l'activité sociale et pédagogique. Du point de vue artistique, le robot fait l'objet d'un projet *Android-Human Theatre*, conçu par Oriza Hirata, projet qui vise à intégrer le Geminoïde F dans un spectacle vivant et à le placer dans une situation de jeu aux côtés des comédiens professionnels. Soulignons que Hirata est aujourd'hui l'un des metteurs en scène et auteurs japonais les plus connus, qui a affirmé sa place surtout grâce à sa « Théorie du style parlé » au théâtre. Il est également le directeur artistique de la troupe Seinendan et mène un travail pédagogique à l'Université d'Osaka. Il possède une connaissance approfondie du théâtre et apporte clairement, dans le projet *Android-Human Theatre*, cette position artistique, sans perdre de vue les aspects sociologiques de la place des robots dans la société japonaise.

Hirata commence à collaborer avec le Laboratoire ATR en 2008, lorsqu'il monte un spectacle, *Je suis un travailleur*, avec deux robots domestiques *wakamaru*. Ensuite, en 2010, il monte *Sayonara* avec Bryerly Long et le Geminoïde F qui marque une étape décisive aussi bien pour l'*Android-Human Theatre* que pour le théâtre du point de vue de son rapprochement avec la robotique et le déplacement de ses propres paradigmes. Hirata place l'androïde non seulement dans une situation de jeu comme il a fait avec le *wakamaru*, mais il lui attribue un rôle scénique égal à celui de la comédienne, les deux devenant désormais partenaires interprètes. Ce choix esthétique trouve sa pertinence dans l'histoire racontée dans le spectacle. Il s'agit d'une jeune femme atteinte d'une maladie incurable en phase terminale. Enfermée dans son petit appartement, elle reçoit de ses parents un androïde qui va lui tenir compagnie. Le spectacle, d'une demi-heure à peine, se joue dans un espace épuré avec deux chaises uniquement et se construit sur un rythme apaisé et dans une ambiance intimiste. Le Geminoïde F est là pour réciter des poèmes, mais en même temps, pour assister la jeune femme dans son dernier départ. L'interaction avec l'actrice réelle se fait au niveau langagier uniquement, la voix du robot est doublée par une autre actrice Minako Inoue qui le téléopère également depuis les coulisses. Il est important de dire que le Géminoïde F n'est pas doté de la capacité de marcher ce qui est inscrit d'une manière tout à fait logique dans la dramaturgie du spectacle.

Indépendamment des questions fondamentales pour le théâtre que pose la démarche de Hirata, nous souhaitons nous arrêter également sur le double contexte de celui-ci : scientifique et artistique. Soulignons que nous avons affaire ici à une situation très intéressante : d'une part, un laboratoire scientifique où le robot a été élaboré et où il est toujours perfectionné et, d'autre part, la scène où il trouve « sa deuxième vie », dans le contexte de la mise en scène et de la théâtralité. Le Laboratoire ATR devient cet espace de travail et de recherche scientifique extrêmement avancé. Il devient également le lieu des essais avec le Geminoïde F qui visent à examiner les solutions technologiques intégrées ainsi que sa pertinence comportementale. Cet aspect a également été décrit par Zaven Paré, l'artiste travaillant avec les robots qui s'est rendu à Osaka et a assisté à plusieurs séries de tests. Il décrit les réflexions autour du déroulement des différents tests dans son ouvrage *Le jour où les robots mangeront des pommes* (coécrit avec Emmanuel Grimaud). Ces témoignages prouvent que l'expérimentation scientifique avec le Geminoïde F transgresse la spécificité du laboratoire technologique en

tant que lieu de recherche, car ce dernier devient l'endroit d'une expérience anthropologique, de la performativité et de la théâtralisation. Le robot, les chercheurs et l'artiste-témoin se trouvent à la fois chercheurs mais également acteurs de l'imprévu, de l'improvisation en dehors du protocole de l'expérience scientifique, de la panne, de l'échec, d'une découverte inattendue. A cette transgression du laboratoire dans sa fonction scientifique s'ajoutent les déplacements intéressants qui s'accomplissent sur le plateau du théâtre. Ce dernier, à son tour, devient non seulement l'espace d'une création artistique mais également cet endroit d'expérimentation avec un robot ultracomplexe. Ici, plusieurs questions se posent sur le plan de la compatibilité des éléments et des logiques scéniques avec le fonctionnement et les contraintes techniques du Géminoïde F. Le laboratoire technologique et le plateau de théâtre dépassent finalement leurs fonctions initiales et redéfinissent la notion de laboratoire tant du point de vue scientifique qu'artistique. Ce dernier devient alors un paradigme intéressant à repenser, car d'une part, il convoque l'idée du laboratoire chère au théâtre, et cela depuis un siècle (Hellerau, Bauhaus, Laboratoire Art et Action) – rappelons seulement que, dans le contexte théâtral, un laboratoire implique un espace clos dédié à un travail approfondi, à une recherche créative où le spectateur est un témoin ponctuel – et, d'autre part, un laboratoire technologique dédié entièrement à la recherche scientifique où le travail progresse d'hypothèse en résultats, et cela à travers de multiples tests soumis à une rigueur et à une pensée analytiques.

Le Géminoïde F constitue ainsi un lien entre les équipes, les différentes fonctions, et permet la rencontre des mondes scientifique et artistique. C'est également une figure qui incite à revoir le paradigme du laboratoire du point de vue de sa mutation récente, due au rapprochement du théâtre et de la technologie. A travers le travail scénique de Hirata, le robot en question s'inscrit dans une expérience très importante consistant en l'attribution d'une fonction actorielle uniquement à un robot humanoïde. Même si Android-Human Theatre est encore une approche fortement expérimentale, Hirata soulève plusieurs interrogations importantes pour la scène et trace ainsi une voie esthétique, sans doute, à suivre.



Fig. 1: « Sayonara » mise en scène : Oriza Hirata, 2010, Phot. Tatsuo Nabu©

References

- Dixon Steve (2007), *Digital Performance. A History of New Media in Theatre, Dance, Performance Art and Installation*, The MIT Press, Cambridge Massachusetts, London..
- Grimaud Emmanuel, Paré Zaven (2011), *Le jour où les robots mangeront des pommes*, Paris, Editions PETRA.
- Pluta Izabella (2012), « *La performance de la machine ou comment les cyborgs et les robots jouent sur la scène* », Ligeia. Dossier sur l'art, N° 117-120, juillet-décembre, pp. 169-185.
- Pluta Izabella (2013), « *Robots sur scène. (En)jeu du futur* », Jeu. Revue de théâtre, N° 149 : Mémoires en jeu, décembre, pp. 145-148.

Arts de la scène et Big Data : conception et développement du logiciel Rekall

de Clarisse Bardiot

Apparu à la fin des années 2000, le Big Data est devenu en moins de temps qu'il n'en faut pour le dire le roi des superlatifs au royaume du numérique. Le marketing, mais aussi l'analyse des risques, ou encore l'épidémiologie sont directement concernés. La culture également. Ainsi, Lev Manovich consacre depuis 2005 ses recherches aux « cultural analytics »[1]. D'après lui, « la numérisation de grands ensembles d'artefacts issus du passé et l'essor des réseaux sociaux dans les années 2000 permettent de renouveler l'étude des processus culturels »[2]. Les études de cas concernent autant l'ensemble des œuvres de Mondrian ou de Rothko que les jeux vidéos ou encore l'évolution graphique des couvertures de Time Magazine depuis sa création. L'enjeu est ni plus ni moins de reconsiderer ce que nous entendons par « culture » ainsi que les méthodologies qui sont appliquées à ce champ.

Au théâtre, le Big Data est officiellement à ses débuts. La première conférence explicite sur le sujet a eu lieu le 9 novembre 2013 dans le cadre du colloque de l'American Society of Theatre Research. Intitulée « Big Data and the Performing Arts », elle est le fait de Doug Reside, conservateur dans le département arts de la scène à la New York Public Library. La conférence concerne essentiellement les archivistes et les chercheurs, lesquels sont confrontés au Big Data des collections et des fonds d'archives numérisés, ou encore aux disques durs des artistes dont ils doivent identifier, archiver et analyser les milliers de données.

La création du logiciel Rekall s'inscrit dans ce contexte et tente de répondre à ces problématiques, ainsi qu'à celles des artistes confrontés à l'obsolescence des technologies numériques. Rekall est un environnement open-source pour documenter, analyser les processus de création et simplifier la reprise des œuvres. Il permet d'agrégier un nombre infini de documents de différentes natures autour d'une création. Ce projet est né dans le champ des *digital performances*, soit des œuvres scéniques à composantes technologiques.

Rekall est né d'une prise de conscience de différents acteurs (artistes, techniciens, programmeurs, chercheurs) concernant les difficultés liées aux technologies numériques employées dans les arts de la scène (notons que face à des problématiques similaires, le champ de la musique contemporaine a tenté d'apporter des réponses, par exemple via les programmes Mustica et Caspar). Les composantes technologiques des *digital performances*, qu'elles interviennent pendant le processus de création (captations vidéo d'improvisations, simulations de mise en scène et de scénographies sur divers logiciels...), ou pendant la représentation (capteurs, dispositifs de téléprésence, images et sons modifiés en temps réel...), renouvellent la question de la documentation des arts de la scène : quelle est la nature des nouveaux documents produits par/pour ces spectacles, comment les analyser, faut-il conserver les programmes informatiques spécifiquement conçus et les rendre accessibles (lisibles) en fonction de l'évolution des programmes et du matériel, dans quelle mesure le hardware et le software doivent-ils être documentés, comment effectuer une captation de ces œuvres ?

Aujourd'hui, toutes les régies techniques sont numériques, et une partie du processus de création a largement lieu via les ordinateurs et les réseaux : échanges de mails, traitements de texte, rendez-vous à distance via des dispositifs de téléprésence (voix sur IP), images et vidéos numériques pour rassembler des idées, des pistes de travail, usage des réseaux de partages d'images pour mettre à disposition des documents visuels pour l'ensemble de la compagnie, croquis effectués sur tablette numérique, etc. Dans ce contexte, l'obsolescence rapide des technologies devient extrêmement problématique, à la fois pour les artistes qui doivent pouvoir continuer à faire tourner leurs spectacles, et pour les chercheurs qui souhaitent en analyser les processus de création. Les documents numériques sont des traces essentielles pour retracer l'histoire des arts de la scène à l'époque contemporaine.

Rekall s'inscrit dans la lignée de recherche menées à l'IRCAM, à la fondation Daniel Langlois, à la Fondation Pina Bausch ou encore par l'équipe de William Forsythe (projet Motion Bank). Sans rentrer dans les détails, il existe différents outils d'annotation vidéo utilisés dans d'autres contextes : Advene, AmiGram, Anvil, ELAN, On the mark, SSI, Vcode/ VData... Pourtant, aujourd'hui, un outil tel que Rekall n'existe pas sur le marché, qu'il s'agisse d'un logiciel gratuit ou payant. Cela n'exclue pas des tentatives qui s'approchent de cette démarche, comme le travail remarquable réalisé par William Forsythe, Maria Palazzi et Norah Zuniga Shaw, mais qui ne s'applique qu'à une seule œuvre.

Rekall est un logiciel qui permet de documenter les *digital performances*, en prenant en compte le processus de création, la réception et les différentes formes d'un spectacle. Il s'adresse à la fois aux artistes, aux techniciens et au grand public. Simple d'usage et rigoureux (en particulier dans les méthodes d'indexation et la gestion des métadonnées), Rekall permet de nombreux usages, au-delà des *digital performances* et des arts de la scène. Pour réaliser ce projet, des structures culturelles (le Phénix scène nationale Valencienne, Le Fresnoy, MA scène nationale Montbéliard), une société (Buzzing Light – Guillaume Marais et Guillaume Jacquemin) ainsi que des institutions (Pictanovo, Ministère de la Culture et de la Communication) se sont associés. Le projet a été initié et conçu par Clarisse Bardiot, en collaboration avec Buzzing Light et Thierry Coduys.

Rekall offre une vision synthétique de la quantité, de la qualité et de l'organisation des documents entre eux, tout en étant au plus près de la démarche propre à chaque artiste, à chaque compagnie. Il permet à la fois de rendre compte des technologies utilisées dans le spectacle et d'en offrir une description pour éventuellement proposer une alternative avec d'autres composantes. Il nous semble en effet primordial de garder la trace la plus précise possible des composantes technologiques, parce qu'elles sont également porteuses de dimensions esthétiques et historiques, tout en offrant la possibilité de décrire les effets de ces mêmes composantes, dans la lignée de la réflexion sur les médias variables.

Le fonctionnement de Rekall s'articule essentiellement autour des documents de création : croquis de scénographies, commentaires audio, description d'éléments techniques, vidéos, textes, carnets de notes, conduites techniques, patches, captures d'écran de logiciels spécifiques, partitions, photographies, mails... Il permet également d'articuler plusieurs strates temporelles : celle du processus de création (éclairer par exemple les recherches menées pour tel aspect du spectacle), de la représentation elle-même (voire de ses différentes versions dans le cas d'un work in progress), et de sa réception (par exemple en ajoutant des commentaires audio de la compagnie sur son propre travail, ou bien de spectateurs, ou encore la revue de presse).

L'accumulation des documents est un élément clé pour l'efficacité de fonctionnement de Rekall. En effet, c'est en analysant ces documents, en les mettant en relation les uns avec les autres et en les plaçant dans des contextes soigneusement choisis (multidimensionnels, temporels ou non) que Rekall parvient peu à peu à révéler la nature de l'œuvre.

Afin de recueillir tous ces documents de travail, il est essentiel que Rekall se trouve au cœur du processus de création. La majeure partie des documents doivent être naturellement implémentés sans représenter une charge de travail supplémentaire pour les artistes, c'est pourquoi une grande partie de Rekall est dédiée à l'organisation des documents de création pendant le processus créatif. Il permet à tous les protagonistes intervenant au cours du processus de création de travailler sur une plateforme commune et compartimentée.

Cette structure ouvre alors un spectre de possibilités analytiques extrêmement important. En partant du principe qu'une œuvre est définie par ses documents de création (devenus documents d'exploitation pour certains), Rekall se base sur les *métadonnées* présentes dans chacun de ces documents pour en extraire des informations cruciales (auteur, date de création, lieu de création, etc.) qui seront ensuite utilisées par les outils d'analyse et de représentation de

l'information, afin de révéler des comportements créatifs, des usages ou d'autres informations insoupçonnées.

Au vu de la masse d'informations que représente une œuvre, il apparaît évident qu'une solution d'export ciblé est nécessaire afin de n'inclure dans différents packages que les documents utiles à l'usage souhaité (pédagogie, exploitation, critique, etc.). Connaissant la nature de chaque document, Rekall permet de configurer simplement ces exports, qui auront également le mérite de valoriser l'œuvre (documents à jour, présentation soignée, etc.).

Rekall est actuellement en version alpha. La bêta est prévue pour mars/avril 2014. La collaboration avec des équipes artistiques dans cette phase d'expérimentation est essentielle. C'est pourquoi nous nous appuyons sur la collaboration de deux équipes artistiques, en théâtre (Jean-François Peyret) et en danse (Mylène Benoit), en résidence au Phénix scène nationale Valenciennes et au Fresnoy. Ces équipes sont d'une part associées à la conception du logiciel et d'autre part les premiers utilisateurs. Des réunions de travail avec les différents intervenants (techniciens, régisseurs, metteur en scène, chorégraphe, éclairagiste, vidéaste) font partie du processus de conception et développement de Rekall, afin d'ajuster régulièrement le cahier des charges et les spécifications aux besoins des futurs utilisateurs.

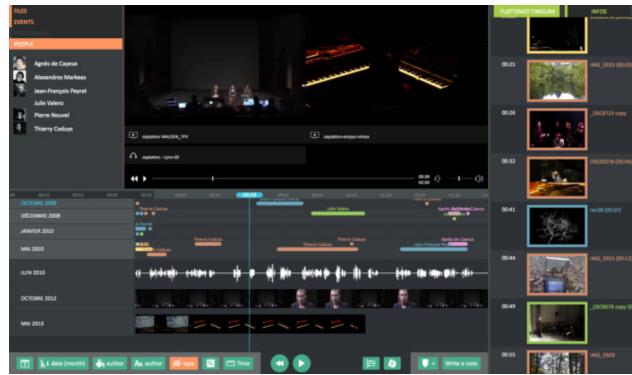


Fig. 2: Rekall - Étude chronologique du processus de création. Etude de cas : Re : Walden, mis en scène par Jean-François Peyret ; axe vertical : année du calendrier ; couleur : type de document ; texte : auteur du document ; axe horizontal : timeline de l'œuvre

References

- International Journal of Performance Arts & Digital Media** (2013), « *Choreographic documentation* », vol. 9, n° 1.
- Performance Research** (2007), « *Digital Resources* », 11:4.
- DeLahunta Scott et Hennemann Célestine** (éd.) (2013), *Motion Bank. Starting Points & Aspirations*, Francfort, Motion Bank / The Forsythe Company.
- DeLahunta Scott et Zuniga Shaw Norah** (2007), « *Constructing memory : Creation of the choreographic resource* », in *Performance Research*, 11:4, pp. 53-62.
- « *Media Visualization: Visual Techniques for Exploring Large Media Collections* », in Kelly Gates (éd.), *Media Studies Futures*, Blackwell, 2012.
- Manovich Lev** (2013), *Software Takes Command*, Bloomsbury Academic.
- Navas Eduardo** (2012), « *Modular Complexity and Remix: The Collapse of Time and Space into Search* », in *AnthroVision* 1.; version revue et corrigée in softwarestudies.com, 4/19/2013.
- Rinehart Richard** (2007), « *A System of Formal Notation for Scoring Works of Digital and Variable Media art* », *Leonardo - Journal of the International Society for the Arts, Sciences and Technology*, The MIT Press, Volume 40, n° 2, pp. 181-187.
- lab.softwarestudies.com
- lab.softwarestudies.com/2008/09/cultural-analytics.html
- A ce sujet, cf. Art Press 2 « *Les enjeux de la conservation des arts technologiques* », n°12, janvier 2009.
- Cf notamment les projets de recherche internationaux Mustica et Caspar. Une partie de ces recherches, notamment sur la conservation des patch, est retranscrite dans ce texte : **Bonardi, Alain et Barthélémy, Jérôme** (2007), « *Le Patch comme document numérique : support de création*

et de constitution de connaissances pour les arts de la performance », in Le Document numérique dans le monde de la science et de la recherche, Actes du 10^{ème} Colloque International sur le Document Numérique (CIDE), INIST, Nancy, p. 168.

Cf. programme de recherche DOCAM sur la documentation et la conservation des arts médiatiques, ainsi que **Bardiot, Clarisse** (2006), *9 Evenings, Theatre & Engineering*, site Internet de la Fondation Daniel Langlois www.fondation-langlois.org/flash/f/index.php?NumPage=571 synchronousobjects.osu.edu

Rethinking Text Reuse as Digital Classicists

Romanello, Matteo

matteo.romanello@dainst.de

German Archaeological Institute & King's College London

Berra, Aurélien

aurelien.berra@u-paris10.fr

Université Paris-Ouest & EHESS

Trachsel, Alexandra

alexandra.trachsel@uni-hamburg.de

University of Hamburg

Rethinking Text Reuse as Digital Classicists

1. Participants

– Panel conveners:

- Aurélien Berra (Université Paris-Ouest & EHESS)
- Matteo Romanello (German Archaeological Institute & King's College London)
- Alexandra Trachsel (University of Hamburg)

– Further participants:

- Monica Berti (University of Leipzig)
- Neil Coffee (University at Buffalo, SUNY)
- Annette Geßner (University of Leipzig)
- Charlotte Tupman (King's College London)

2. Description of the panel

2.1 Why rethink text reuse

Text reuse is the meaningful reiteration of text, usually beyond the simple repetition of common language. Such a broad concept can naturally be understood at different levels and studied in a large variety of contexts. This diversity of approaches is to some extent explained by the fact that the phenomenon exists in almost all disciplines of the Humanities, and is crucial in those which focus on texts.

At one end of this spectrum we find the methods developed by computational linguistics. Research projects in this field study text reuse through automatic analyses within large corpora that often come from widely different backgrounds. The approaches range from the automatic detection of allusions and intertextual phenomena, for example in historical texts, to the detection of plagiarism in modern ones^{1 2 3 4}. At the other end, the concept also designates a core scholarly activity, connected to most of the “scholarly primitives”⁵ – this meta-level being obviously our own practice, and having its roots in Antiquity. Furthermore, any kind of citation constitutes an indirect way of transmitting knowledge, either consciously or unconsciously, as well as a rhetorical or narrative device allowing an author to communicate with his audience beyond the level of the linguistic

content. As a result, this notion shows how deeply intertwined objectivity and subjectivity are when one handles texts.

Digital approaches often aim at highlighting or defining these complex links between an initial statement and its multiple occurrences (often translations) in later contexts. Indeed, especially when the text reuse of ancient elements in corpora of more recent texts is studied, the fact that the statements are given in translation is an important issue and introduces an additional difficulty. This, however, is not a completely new problem. It can be observed each time that two cultures meet and borrow elements from each others' cultural heritage. A further notion of text reuse is reached when not only the interconnections between the different reuses of a given textual element are investigated, but also the connections between the contexts in which they occur, whether in the form of unabbreviated quotations or as references within a more conventional citation system.

This panel proposes to gather researchers from different projects focusing on text reuse in order to create an inventory of the possible approaches to and understandings of the notion. Our objective is to highlight the historical dimension of the phenomenon and, ultimately, find some common features that could lead to a more systematic study. Texts are data indeed, but text reuse provides an excellent demonstration that they must be studied also and at the same time as intentional, sophisticated and reflexive cultural products. The emergence of Digital Classics, and of Digital Humanities in general, is an occasion to rethink text reuse and work towards the integration of – or at least foster dialogue and interconnection between – various perspectives.

2.2 Studying text reuse in Digital Classics

A panel on text reuse at the Digital Humanities 2014 Conference seems a very timely initiative, because several projects are currently addressing the question and developing new tools to deal with its different aspects.

The Perseids platform^{6 7} can be mentioned first. As a project of the Perseus Digital Library⁸, it aims at creating a collaborative online environment for the edition of a great variety of ancient documents, privileging the requirements of the editing of fragmentarily preserved sources (especially if they are transmitted through quotations) – a specific case of text reuse^{9 10 11 12}. Indeed current digital libraries, like the Perseus Digital Library or the Thesaurus Linguae Graecae, have started with the wholly preserved ancient texts and deal with fragmentary works as if they were independent entities at the same level as the others. This clearly creates conceptual difficulties, since we only have indirect access to most of the fragmentarily preserved work: some parts of a lost initial work have been reused in the form of quotations in later texts. This reuse may have left some traces in the rewording of the quotation and therefore it is essential to keep the link to the context in which a given passage has been embedded when editing fragmentarily preserved texts.

One way of addressing this issue has been explored by the Sharing Ancient Wisdoms project¹³. The project's goal was to provide digital editions of several texts belonging to the so-called tradition of wisdom literature, by analysing the quoted sayings or proverbs and creating an ontology allowing to describe their diverse relationships¹⁴. Still another approach must be chosen when the focus is shifted from the edition of a text with many quotations in it, such as those dealt with in the SAWS project, to the edition of a set of quotations that come from different source texts, but belong to one lost work, as is currently being explored in Alexandra Trachsel's research on Demetrios of Scopis.

In a complementary fashion, the study of single works of considerable size as webs of quotations should enable us to deal better with the reflexive dimension of encyclopaedic writings. Such a perspective is being built in the Digital Athenaeus project, which will explore the combination of digital and philological means of analysis in the preparation of a new edition of the Deipnosophists – a complex literary construction which sets scholarly discussions and pastimes in the context of

an Imperial symposium and thus witnesses to the dynamics of text reuse¹⁵.

Further projects, such as *Tesserae*¹⁶ or *Eumaios*¹⁷ move beyond the concept of quotation and focus on more hidden or less acknowledged forms of intertextuality. *Tesserae*, in particular, is devised to help scholars find previously unexplored intertextual parallels by means of automatic text reuse detection¹⁸. This work has employed small benchmark sets of recognised parallels against which search techniques are measured and methods are improved. But having at hand a large and systematic repertory of already studied *loci parallelii* is something from which a tool like *Tesserae* will benefit immensely and that can be built, to a large extent automatically, by extracting from the literature the text passages that were already studied in relation to one another. These parallels are usually signalled in journal articles and other types of secondary sources by means of canonical citations, whose automatic extraction from large corpora of unstructured texts, such as those of JSTOR or the Internet Archive, is a topic that is currently being explored^{19 20}.

The identification and extraction of text reuse is central in eTRACES, a project which just developed a tool named GERTRUDE (Göttingen E-Research Text-Re-Use for Digital Editions). Working on extremely heterogeneous corpora and primarily on German literature written between 1500 and 1900, it actually reflects on and solves similar problems.

All these projects, though they have the concept of text reuse in common, can be distinguished either by the type of corpus they use (texts from Antiquity, German literature, modern scholarly writings) or by their starting point (working on source texts where quotations are preserved, establishing relationships between different works in which the same textual elements occur, or focusing on quoted or reused elements). However, they have accumulated a great amount of knowledge on how to deal with the multiple forms of this cultural practice. The panel therefore aims at bringing together these efforts and should allow each of the projects to benefit from the expertise of the others, so that the solutions already found may be discussed and in the hope that our desiderata may lay the ground for further research.

3. Practical organisation of the panel

Besides the conveners, who will introduce and moderate the discussion, the panel will involve four speakers. After a brief presentation of the participants and of the main issues of the topic (10 minutes), short talks will be given by the four panel participants, illustrating different aspects of text reuse (40 minutes). The remaining time will be devoted to a discussion among all the participants and will be focused on the challenges and desiderata for further projects dealing with text reuse, in Digital Classics and beyond this field (40 minutes).

References

1. **Bamman, D., & Crane, G.** (2008). *The logic and discovery of textual allusion*. In: Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008), Marrakesh.
2. **Büchler, M.** (2013). *Informationstechnische Aspekte des Historical Text Re-use*. PhD Thesis, Universität Leipzig. Retrieved from nbn-resolving.de/urn:nbn:de:bsz:15-qucosa-108515.
3. **Bamman, D. & Crane, G.** (2009). *Discovering Multilingual Text Reuse in Literary Texts*. Available at www.perseus.tufts.edu/publications/2009-Bamman.pdf
4. **Lee, J.** (2007). *A Computational Model of Text Reuse in Ancient Literary Texts*. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pp. 472–479. Prague, Czech Republic: Association for Computational Linguistics. Retrieved from acl.ldc.upenn.edu/P07/P07-1060.pdf .
5. **Unsworth, J.** (2000). *Scholarly Primitives: What Methods Do Humanities Researchers Have in Common, and How Might Our Tools Reflect This? Formal methods, experimental practice*. King's College, London. people.lis.illinois.edu/~unsworth/Kings.5-00/prim
6. **Almas, B. & Berti, M.** (2013). *Perseids Collaborative Platform for Annotating Text Re-Uses of Fragmentary Authors*. In F. Tomasi & F. Vitali, eds. DH-Case 2013. Available at dx.doi.org/10.1145/2517978.2517986.
7. **Perseids**. *A collaborative editing platform for source documents in Classics*, sites.tufts.edu/perseids
8. *Perseus Digital Library*, www.perseus.tufts.edu (Accessed on November 1, 2013).
9. **Romanello, M., Berti, M., Boschetti, F., Babeu, A., & Crane, G.** (2009). *Rethinking Critical Editions of Fragmentary Texts by Ontologies*. In S. Mornati, ed., Rethinking Electronic Publishing: Innovation in Communication Paradigms and Technologies - Proceedings of the 13th International Conference on Electronic Publishing held in Milano, Italy 10-12 June 2009, pp. 155-174.
10. **Romanello, M.** (2011). *The Digital Critical Edition of Fragments: Theoretical Problems and Technical Solution*. In P. Kurras Cotticelli, ed., *Linguistica e Filologia Digitale: Aspetti e Progetti*, pp. 147–155. Alessandria: Edizioni dell'Orso. Retrieved from eprints.rclis.org/handle/10760/15592 .
11. **Trachsel, A.** (2012). *Collecting Fragments Today: What Status Will a Fragment Have in the Era of Digital Philology?* In C. Clivaz, J. Meizoz, F. Vallotton, & J. Verheyden, eds., *Lire demain – Reading Tomorrow*, pp. 415-429 (ebook). Lausanne: Presses polytechniques et universitaires romandes.
12. **Berti, M.** (2013). *Collecting Quotations by Topic: Degrees of Preservation and Transtextual Relations among Genres*. Ancient Society, 43, pp. 269–288.
13. *Sharing Ancient Wisdoms*, www.ancientwisdoms.ac.uk (Accessed on November 1, 2013).
14. **Dunn, S., Hedges, M., Jordanous, A., Lawrence, K. F., Roueché, C., Tupman, C. & Wakeling E** (2012). *Sharing Ancient Wisdoms: Developing Structures for Tracking Cultural Dynamics by Linking Moral and Philosophical Anthologies with their Source and Recipient Texts*. In Digital Humanities Conference, Hamburg, Germany. Available in the Book of Abstracts at www.dh2012.uni-hamburg.de/conference/programme/abstracts, pp. 176-179.
15. **Romanello, M., & Berra, A.** (2011). *The Critical Step in Open Content Greek: Towards a Digital Edition of Athenaeus*. In TEI Members Meeting, Würzburg, Germany. Available in the Book of Abstracts at www.zde.uni-wuerzburg.de/tei_mm_2011, pp. 43-47, and philologia.hypotheses.org/512 .
16. *Tesserae*, tesserae.caset.buffalo.edu (Accessed on November 1, 2013).
17. *Eumaios: a collaborative website for Early Greek epic*, panini.northwestern.edu/AraServer?eumaios+0+frame.anv (Accessed on November 1, 2013)
18. **Coffee, N., Koenig, J.-P., Poornima, S., Forstall, C. W., Ossewaarde, R., & Jacobson, S. L.** (2013). *The Tesserae Project: Intertextual Analysis of Latin Poetry*. Literary and Linguistic Computing, 28(2), pp. 221–228. DOI: 10.1093/linc/fqs033.
19. **Romanello, M.** (2013). *Creating an Annotated Corpus for Extracting Canonical Citations from Classics-Related Texts by Using Active Annotation*. In A. Gelbukh, ed., Computational Linguistics and Intelligent Text Processing. 14th International Conference, CICLING 2013, Samos, Greece, March 24-30, 2013, Proceedings, Part I. Springer, Berlin Heidelberg, pp. 60–76.
20. **Romanello, M., Boschetti, F. & Crane, G.** (2009). *Citations in the Digital Library of Classics: Extracting Canonical References by Using Conditional Random Fields*. In Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries. Morristown, NJ, USA: Association for Computational Linguistics, pp. 80–87.
21. *eTRACES*, etraces.e-humanities.net (Accessed on November 1, 2013).

What is Modeling and What is Not?

Van Zundert, Joris

joris.van.zundert@huygens.knaw.nl

Huygens Institute for the History of the Netherlands, Royal Netherlands Academy of Arts and Sciences

Jannidis, Fotis

fotis.jannidis@uni-wuerzburg.de
Würzburg University

Drucker, Johanna

drucker@gseis.ucla.edu
University of California, Los Angeles

Rockwell, Geoffrey

grockwei@ualberta.ca
University of Alberta

Underwood, Ted

tunder@illinois.edu
University of Illinois, Urbana-Champaign

Kestemont, Mike

mike.kestemont@gmail.com
Antwerp University

Andrews, Tara

tara.andrews@kps.unibe.ch
Bern University

What is Modeling and What is Not?

In A Companion to Digital Humanities¹, Willard McCarty cites Nelson Goodman in saying that the term 'model' can be used to denote "almost anything from a naked blonde to a quadratic equation". Indeed the terms 'model' and 'modeling' seem almost painfully polysemous. Nevertheless within Digital Humanities we cannot ignore the terms or the concepts behind them—the notions are inextricably linked to what is one of the core objectives of humanities computing², namely to render humanities data computationally tractable³ and processable^{4 5} to enhance our abilities for analysis.

In light of the renewed debate on modeling in Digital Humanities⁶ this panel proposes to investigate how humanists currently understand the role and meaning of modeling, and how we may arrive at an understanding of the term appropriate for humanities research and pedagogy.

McCarty stated a decade ago that the humanities lack a disciplined way of talking about modeling⁷ which makes it extremely difficult to define the properties and uses of appropriate models for humanities research. Modeling is a commonplace implicit activity in digital humanities, yet our modeling activities are almost never explicitly discussed as such, and it is rarely pointed out that many of our results are in fact models: charts, probabilistic methods, interfaces to the information we structure in databases. This implicitness is attested by our language use. We do not speak of "modeling an analogy" or of "modeling a chart". We "make" or "create" them as concrete representations of an implicit and abstract model.

Yet, given the concrete applications and results that can already be seen within the humanities, modeling needs to be a humanities praxis to the same extent as it already is in other scientific fields such as biology and physics. As the social sciences –more specifically the ethnography practices in Science & Technology Studies for instance– show us, praxis by definition can be studied and interrogated for its properties by observing and following its practitioners⁸. This panel provides a first step in such observant interrogation.

In the computational domain modeling can be delineated in a narrow mathematical sense where model theory⁹ defines Turing complete languages as models or instantiations of logic constructed from formulas (i.e. syntax or rules) and signatures (i.e. vocabulary or objects). Thus, computer languages are themselves mathematical models of logic. They provide a layer of expressive logic that in turn allows us to compositionally model data, objects and their relations¹⁰. Analogous to the statement made by Peter Robinson about interfaces¹¹, we can argue that such a composition or model expresses an intellectual argument about the real world entities and relations

they mimic, capture, or simulate—an intellectual argument that is made on several levels through the computational model and that eventually is communicated to an observer (or user) by way of its interface.

In recent years we find most notably the application of modeling in order to create maps, graphs, trees^{12 13}, analogies, diagrams, charts, simulations¹⁴, and stylometric analyses¹⁵, as well as in discourse analysis, topic modeling, and narrative modeling¹⁶. If the successful computational analytical models are quantitatively and statistically founded, does that mean that humanities modeling must necessarily be anchored in the somewhat narrowly defined models that are generally associated with quantification and computer science

More generally, must the concepts of 'model' and 'modeling' appropriate for Digital Humanities be bound solely by parameters of the mathematical foundations of binary logic? Modeling as activity and concept applies more widely to the humanities than merely in its computational applications. Is it possible to turn around the dynamic of the computational 'stack', so that rather than having mathematics drive humanities computability, the properties of humanistic problems and the data behind them might drive models of computation? We can argue that the goal of any computational approach within the humanities is to render computable the complexity, the abstraction, the ambiguity, the subjectivity, and multiplicity of perspective of the humanities¹⁷. Similarly: how do we encompass aspects of modeling present in simulation and (serious) gaming¹⁸ of which the humanistic aspects seem to transcend the narrow mathematical connotation of 'model'? And how does modeling relate to the continuing history of developing and redeveloping digital humanities tools that—rather than merely representing infrastructure—creates a record of intellectual theorizing humanistic computational models?¹⁹ How do we break out of the mathematical sandbox defined by first-order logic to do justice to the modalities of humanities? Does this require completely new models for data, logic, and representation? Does it require a general theory of modeling? Even a new symbolic language inspired by the humanities?

This panel brings together some of the most visible practitioners of computational methods within the humanities who have captured analytic models in software code, as well as some of the most influential figures of what might be called 'tacit modeling theory in digital humanities'. We invite them to consider the characteristics of humanities modeling and how those contrast with computational modeling and mathematical modeling, so as to determine what idiosyncrasies modeling might have in a humanities domain. Do these idiosyncrasies allow us to delineate a computationally tractable vocabulary at all? To investigate these questions the panel will discuss and reflect on matters such as...

- How do we address the role of modeling and models in the humanities?
- How do we ensure that existing mathematical logic does not confine our ability to represent and manipulate humanistic evidence?
- What benefits does a definition of modeling appropriated for the humanities hold?
- What would a symbolic language for the humanities look like?
- What are the standards of evaluation in modeling and do we need specific ones in the Humanities?
- What is a useful vocabulary to talk about modeling in a humanities sense?

Panelists

Joris van Zundert (Chair) is in charge of methodological research at the Huygens Institute for the History of the Netherlands. Next to his research in computational humanities he is interested in exchanges between digital humanities and science and technology studies (STS).

Tara L. Andrews has implemented a digital workbench for the fully computational stemmatic analysis of text traditions (<http://www.digitalbyzantinist.org/2012/09/announcing-stemmaweb.html>). As an assistant professor of digital humanities she is currently developing and teaching a

curriculum that emphasizes modeling and algorithmic approaches to humanistic analysis

Johanna Drucker vehemently called attention to the properties of humanities data that are normally neglected by mathematical and conventional computational models and analyses. She has argued that all data are in fact *capta* and that naïve approaches to statistics are at risk of defining all data as intrinsically quantitative

Fotis Jannidis is developing a white paper on modeling in digital humanities, a version of which will be included in the new edition of the Companion to Digital Humanities. He is a member of the TEI consortium—most notably as the Chair of the Genetic Edition Encoding Special Interest Group. TEI can be designated the only de facto standard for text structure modeling and encoding

Mike Kestemont specializes in stylometry and together with the Computational Stylistics Group (<https://sites.google.com/site/computationalstylistics/home>) has developed "Stylo", a software package in the R statistical programming language. He is an expert of statistical models expressed through computer algorithms and applied to literature stud

Geoffrey Rockwell conceptualized a number of highly visible tools for text analyses (e.g. Voyant: <http://voyant-tools.org>). He is finalizing a book demonstrating amongst others how the hermeneutic and theoretical aspects of text analysis models in the form of tool development transcends mere IT mathematics and infrastructure.

Michael Sperberg-McQueen is a markup specialist by profession and was co-editor of the Extensible Markup Language (XML) specification, chair of the XML Schema working group, as well as heavily involved with the Text Encoding Initiative (TEI)

Ted Underwood works at the interface of literary history and machine learning and is particularly interested in using Bayesian statistics to develop models that reason about uncertainty in a principled way. He maintains an influential blog on his experiences in computational humanities (<http://tedunderwood.com/>).

Organization of the panel

The primary selection criterion for the panelists is their expertise, but care has been taken to balance the panel as much as possible for age, gender, field, and region. The panel session will be organized as follows

- The Chair will introduce the panel's topic, discussion questions, and the panelists (10 minutes);
- Each of the panelists will give a definition of modeling as a 1 minute provocative pitch (10 minutes);
- An open forum between the panelists and the audience follows (60 minutes);
- A circular setting of seats with panelists distributed among the attendees will be used to enhance audience participation in the discussion;
- The panel discussion will be audio recorded, concise conclusions will be published to the web.

Further Reading

- Checkland, P. & Holwell, S., 1998. *Information, Systems, and Information Systems: Making Sense of the Field*.** Chichester: John Wiley & Sons, Ltd.
- Davis, M., 2012. *The Universal Computer: The Road From Leibniz to Turing*.** New York: CRC Press.
- Mahoney, M.S., 2011. *Histories of Computing*.** T. Haigh (ed.), Cambridge: Harvard University Press.
- Hayles, K.N., 2012. *How We Think: Digital Media and Contemporary Technogenesis*.** Chicago: University of Chicago Press.
- Ramsay, Stephen, 2011. *Reading Machines: Toward an Algorithmic Criticism (Topics in the Digital Humanities)*.** Chicago: University of Illinois Press.

1. Schreibman, Susan, Raymond George Siemens, and John M. Unsworth (2004). *A Companion to Digital Humanities*. Wiley-Blackwell.

2. Unsworth, J., (2002). *What Is Humanities Computing And What Is Not?* G. Braungart, P. Gendolla, & F. Jannidis, eds. *Jahrbuch für Computerphilologie*, 4. Available at: computerphilologie.digital-humanities.de/jg02/unsworth.html (Accessed July 8, 2013).

3. McCarty, W. (2005). *Humanities Computing*, New York: Palgrave MacMillan.

4. Unsworth, J., (2002). *What Is Humanities Computing And What Is Not?* G. Braungart, P. Gendolla, & F. Jannidis, eds. *Jahrbuch für Computerphilologie*, 4. Available at: computerphilologie.digital-humanities.de/jg02/unsworth.html (Accessed July 8, 2013).

5. Orlandi, T., *The Scholarly Environment of Humanities Computing, A Reaction to Willard McCarty's talk on The computational transformation of the humanities*. Available at: rmcisadu.let.uniroma1.it/~orlandi/mccarty1.html (Accessed May 7, 2012).

6. Flanders, J. & Jannidis, F., (2012). *Panel Discussion: Data Models in Humanities Theory and Practice*, Providence (US). Available at: youtu.be/lHJmPT-VjPE (Accessed November 1, 2013).

7. McCarty, W., (2004). *Modeling: A Study in Words and Meanings*. In S. Schreibman, R. Siemens, & J. Unsworth, eds. *A Companion to Digital Humanities*. Oxford: Blackwell. Available at: www.digitalhumanities.org/companion/.

8. Kaptelinlin, V. & Nardi, B.A., (2006). *Acting with technology: activity theory and interaction design*, Cambridge, MA, USA/London UK: MIT Press.

9. Rautenberg, W., (2009). *A Concise Introduction to Mathematical Logic* 3rd ed., Available at: page.mi.fu-berlin.de/raut/logic3/announce.pdf.

10. Forbus, K.D., (2008). *Qualitative Modeling*. In F. van Harmelen, V. Lifschitz, & B. Porter, eds. *Handbook of Knowledge Representation. Foundations of Artificial Intelligence*. Amsterdam, Boston, Heidelberg etc.: Elsevier, pp. 361–394.

11. Robinson, P., (2013). *Five desiderata for scholarly editions in digital form*. In Digital Humanities Conference 2013. Lincoln (NB, USA). Available at: dh2013.unl.edu/abstracts/ab-314.html.

12. Moretti, F. (2007). *Maps, Graphs, and Trees: Abstract Models for Literary History*. London: Verso.

13. Jockers, M. (2013). *Macroanalysis: Digital Methods and Literary History*. University of Illinois Press.

14. McCarty, W. (2005). *Humanities Computing*, New York: Palgrave MacMillan.

15. Hoover, D.L., (2012). *The Excel Text-Analysis Page: A Collection of Microsoft Excel © spreadsheets with macros, in the service of text-analysis*. Available at: files.nyu.edu/dh3/public/The%20Excel%20Text-Analysis%20Pages.html (Accessed October 14, 2013).

16. Meister, J.C. & Gertz, M., (2013). *heureCLÉA, collaborative literature exploration & annotation*. heureCLÉA | Tools. Available at: heureclea.de/tools/.

17. Drucker, J., (2011). *Humanities Approaches to Graphical Display*. *Digital Humanities Quarterly*, 5(1). Available at: digitalhumanities.org/dhq/vol/5/1/000091/000091.html (Accessed August 24, 2012).

18. Bogdanovych, A., Cohen, A. & Roper, M., (2009). *The City of Uruk: Virtual Institutions in Cultural Heritage*. In Proceedings of the HCSNet 2009 Workshop on Interacting with Intelligent Virtual Characters. HCSNet 2009 Workshop on Interacting with Intelligent Virtual Characters. Sydney. Available at: www-staff.it.uts.edu.au/~anton/Publications/HCSNet09.pdf.

19. Ramsay, S. & Rockwell, G., (2012). *Developing Things: Notes toward an Epistemology of Building in the Digital Humanities*. In Debates in Digital Humanities. University of Minnesota Press. Available at: dhdebates.gc.cuny.edu/debates/text/11.

Papers

"Civilization arranged in chronological strata": A digital approach to the English semantic space

Alexander, Marc

marc.alexander@glasgow.ac.uk
University of Glasgow

Anderson, Wendy

wendy.anderson@glasgow.ac.uk
University of Glasgow

1. Introduction

This paper focuses on the history of the English lexicon, and on displaying a new approach to this history through the database of the *Historical Thesaurus of English*¹ (hereafter abbreviated to HT). It does so by reference to the semantic space of English, following Lehrer's statement that 'the words of a language can be classified into sets which [...] divide up the semantic space or the semantic domain in certain ways'. This space is described in this paper as the total accumulation of the various individual semantic fields which make up the language, as represented in the HT database. The paper therefore computationally analyses the size of the English lexicon in these semantic clusters over time, including the metaphorical links which weave between these fields, and so aims to demonstrate the use of the HT in digital humanities by giving a digital analysis of the history of English in ways which were previously not possible.

As part of two wider projects,^{2 3} the present paper focuses on describing the general empirical outlines and development of the English semantic space, accompanied by a case study of three contrasting semantic fields and their metaphorical relationships. These are outlined below, following a description of the methodology and theoretical basis of the paper.

2. The *Historical Thesaurus* and Lexical History

The data used in this paper is drawn from the database of the HT, which arranges into hierarchical semantic categories all the recorded words expressed in English from Anglo-Saxon times to the present day, with 793,742 entries within 225,131 categories, each category representing a distinct concept. These concepts are arranged hierarchically and semantically, so that each concept is placed near or within other, similar concepts.

In so doing, the HT unlocks the linguistic and historical data which is currently inaccessible in any usefully-structured way inside historical dictionaries such as the *Oxford English Dictionary* (OED)⁴. As Charlotte Brewer says, with reference to a review in the *Times of the OED*:

"...even the intensively habitual user [of the OED] could not hope to construct, from an overwhelming multiplicity of individual items, the complete picture, 'the various forms of [...] civilization arranged in chronological strata'..."⁵

Alphabetical arrangement, absent any alternative structure, makes this construction incredibly difficult, if not impossible. But the HT, which structures itself based on meaning and not the alphabet, does give researchers access to this 'complete picture'. This was one of the intentions of the HT from the beginning: Professor Michael Samuels, founder of the project in 1964, saw it as a way of revealing the information about social and cultural change inside and throughout the lexicon which was not easily available for researchers to access.⁶ The HT is therefore a massive digital resource for the study of this phenomenon.

3. Semantic Space

Key for the first part of this paper is that the HT, when analysed in database form, gives an indication of rates of lexicalisation in the history of English. This relates to the phenomenon of synonymy, a situation in a language where a number of words are created (or lexicalized) for a single concept (for more on the following discussion, see, amongst others, Lyons 1995⁷, Verhagen 2007⁸, Hughes 1989⁹, and Taylor 2003¹⁰). While synonymy is a common occurrence in English, as in many other languages, the linguistic insight that synonymy is a form of recategorization, where speakers create a new synonymous term because they wish to reflect a shift in their understanding of, or attitudes towards, a particular concept, allows the use of data on lexicalization rates as an indicator of particular speaker attention to a given concept. Therefore, a situation where there are multiple words for a given concept reflects the evolution of speakers' reactions, attitudes, perceptions and awarenesses of that concept, as human language is too efficient a system to permit there to exist large sets of undifferentiated terms which mean precisely the same thing. The present paper therefore uses this measure as a rough proxy for importance of a concept (just as frequency is used as a similar measure of importance in corpus linguistics, with all the associated issues that varying corpus construction techniques brings with it).

Therefore, for the first time, the database of the HT can give us an empirically based view of English by viewing the changes in the internal structure of the language from an entirely semantic viewpoint. By separating the story of English into the multiple stories of interacting and interrelated semantic fields, this approach can describe the history of English as one generally characterized by overall growth accompanied by occasional trauma which results in sudden expansions or contractions of the English lexicon. The rate of change of each semantic field is therefore a statistic which demonstrates the incidence of such instances of trauma, growing or declining in response to external and internal factors either particular to a semantic field or general to the language as a whole. Such general factors in English include the well-known sudden growth in the mid-1400s which occurs at the start of the English Renaissance, and the Elizabethan and Jacobean spurt which begins a little after 1550, which can be seen in Figure 1:

The Growth of English 1050-2000

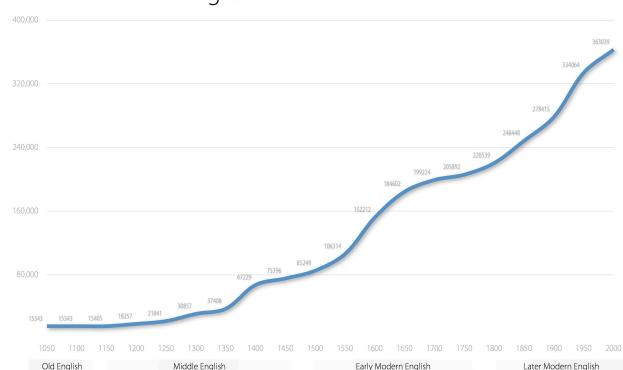


Fig. 1: The growth of the English language across time, as recorded in the HT.

In addition to presenting an overview of the growth of the semantic space of English between the years 1100 and 2000, the paper will also give short case studies of three aggregate semantic fields (figure 2) and their metaphorical relationships (see section 4 below):

- 02.01.15 *Attention and Judgement*, a very large and highly variable category, with an increase of 1000 words in the 1575-1600 period, but a fall of 261 in 1875-1900.
- 03.10.13 *Trade and Commerce*, a category which is relatively small but has one of the highest rates of growth spurts, punctuated with long plateaus.
- 03.05.05 *Moral Evil*, a category which peaks in 1650 and is one of the rare examples of frequent decline across the

history of English, with a loss of 246 words between 1650 and 1900.

The Growth of Three Semantic Categories 1050-2000

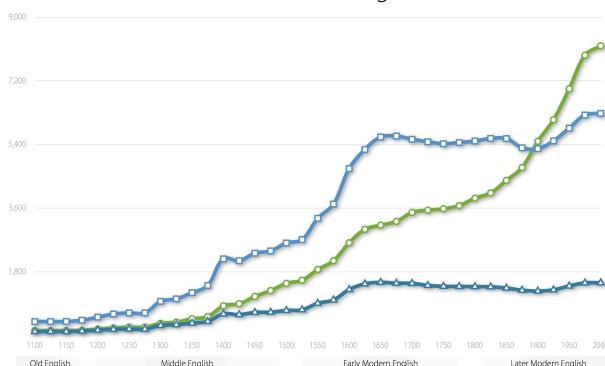


Fig. 2: The growth of three semantic fields. Square: 02.01.15 Attention and Judgement; Circle: 03.05.05 Moral Evil; Triangle: 03.10.13 Trade and Commerce.

Each of these reflect both global trends in the history of English (such as those above, in addition to relative plateaus in the 1700s) while also containing their own internal factors, such as shifts in religious emphasis and in broader economic and industrial patterns.

Not all of these factors are expected; there is no mention in the literature of the rise and fall of lexicalization in the semantic field of *Moral Evil*, nor in many of the other unusual patterns in the rate-of-change data described in this paper. The new data described here gives rise, in the tradition of digital humanities, to the necessity for further explanations from a range of humanities disciplines, such as linguistics, history and literary studies (see Alexander and Struan 2013¹¹ for an interdisciplinary study in a further semantic field).

4. Metaphoricity

Beyond these rates of change, each semantic field above has metaphorical links to other areas of the language, which the HT can reveal to us. Far from being a solely literary technique, much of all language is figurative – recent research has shown somewhere between 8% and 18% of English discourse is metaphorical, with an average of every seventh word being a metaphor.¹²

This is problematic, as while advances are being made in the semantics of digital texts, alongside emerging concepts of a semantically-aware Web, we are at a very early stage in comprehensively and systematically understanding English metaphor, and therefore at an early stage of being able to accurately deal digitally with the meanings encoded in those texts. By mapping the HT's semantic categories onto one another in order to analyse the degree of lexical overlap in different conceptual fields, we can provide results which will comprehensively demonstrate the widespread, systematic and far-reaching impact of metaphor on English. This is the aim of the *Mapping Metaphor* project at Glasgow,¹³ which provides some of our data in this paper, demonstrating empirically the systematic lexical connections between our case study fields (such as that between attention and vision, or evil and darkness).

5. Conclusion

Overall, as well as giving an overview of the history of the English semantic space and its metaphorical interrelationships, the paper also argues for a semantically-informed history of English which operates from a top-down approach, picking out broad patterns and the connections between various semantic categories in order to highlight for analysis those noteworthy elements in a large sea of data. As ever, such large-scale

analyses are only possible through a combination of database techniques, statistical analysis, visual displays of complex datasets, and humanities scholarship.

References

1. Kay, C., J. Roberts, M. Samuels, and I. Wotherspoon (eds). (2009). *Historical Thesaurus of the Oxford English Dictionary*. Oxford: Oxford University Press.
2. Alexander, M. (2012). *Patchworks and Field-Boundaries: Visualising the history of English*. Conference paper at Digital Humanities 2012. Hamburg: University of Hamburg.
3. Anderson, W., M. Alexander, E. Bramwell, C. Kay, and C. Hough. (2013)–. *Mapping Metaphor with the Historical Thesaurus*. Glasgow: University of Glasgow. www.glasgow.ac.uk/metaphor
4. Simpson, J. and E. Weiner (eds). 1989. *The Oxford English Dictionary*, 2nd edition. Oxford: Oxford University Press.
5. Brewer, C. (2007). *Treasure-House of the Language: The Living OED*. New Haven, CT: Yale University Press. Page 232.
6. Samuels, M.L. (1972). *Linguistic Evolution: With Special Reference to English*. Cambridge: Cambridge University Press. Page 180.
7. Lyons, J. (1995). *Linguistic Semantics*. Cambridge: Cambridge University Press.
8. Verhagen, A. (2007). *Construal and Perspectivization*. In The Oxford Handbook of Cognitive Linguistics, eds D. Geeraerts and H. Cuyckens. Oxford: Oxford University Press. 48-81.
9. Hughes, G. (1989). *Words in Time: A social history of the English vocabulary*. Oxford: Basil Blackwell.
10. Taylor, J.R. (2003). *Linguistic Categorisation*, 3rd edition. Oxford: Oxford University Press.
11. Alexander, M and A. Struan. (2013). 'In countries so unciviliz'd as those?': *Notions of Civility in the British Experience of the World*. In Experiencing Imperialism, eds M. Farr and X. Guégan. London: Palgrave Macmillan.
12. www.glasgow.ac.uk/metaphor

Metaphor, Popular Science and Semantic Tagging: Distant Reading with the Historical Thesaurus of English

Alexander, Marc

marc.alexander@glasgow.ac.uk
University of Glasgow

Anderson, Jean

jean.anderson@glasgow.ac.uk
University of Glasgow

Baron, Alistair

a.baron@lancaster.ac.uk
Lancaster University

Dallachy, Fraser

fraser.dallachy@glasgow.ac.uk
University of Glasgow

Kay, Christian

christian.kay@glasgow.ac.uk
University of Glasgow

Piao, Scott

s.piao@lancaster.ac.uk
Lancaster University

Rayson, Paul

p.rayson@lancaster.ac.uk
Lancaster University

1. Introduction

This paper describes and implements a computational procedure for semantically analysing analogy in large bodies of text using a semantic annotation system based on the database of the *Historical Thesaurus of English*.¹ In so doing, it demonstrates the value of a comprehensive and fine-grained semantic annotation system for English within corpus linguistics. Using log-likelihood measures on its semantically-annotated corpus of abstract popular science, the paper therefore demonstrates the existence, the extent, and the location of significant metaphorical content in this corpus. In so doing, it applies a version of Franco Moretti's 'distant reading' programme in the analysis of literary history to non-narrative texts, as well as continuing work on integrating meaning into the methodologies of corpus linguistics.²

1.1. Analogy and Popular Science

Following the 1980 publication of George Lakoff and Mark Johnson's *Metaphors We Live By*,³ it has been frequently stated that human beings, as embodied minds perceiving the mental, social and physical worlds around them, understand abstractions in terms of concrete entities. While this is a well-explicated concept in cognitive linguistics and psychology, few studies have yet aimed to establish both the extent and operation of this in a large corpus of discourse. The standard methodology in cognitive linguistics tends to rely on introspection and the intuitions of native speakers, at the expense of empirical data.⁴ This lack of rigour has resulted in results which, though "intuitively appealing", are criticized "for lacking a clear set of methodological decision principles".⁵ Following earlier work we have undertaken on the investigation of analogy and metaphor in English from empirical groundings,⁶ in this paper we discuss a methodology for identifying these textual phenomena automatically, and in so doing aim to open up cognitive linguistics to more digital humanities techniques, in addition to demonstrating the use of automated semantic annotation and disambiguation techniques at an unprecedented level of granularity.

1.2. The Corpus

We take as our initial data two book-length popular science texts which focus on explaining abstract concepts to a non-specialist audience, and therefore provide the greatest potential for the analysis of non-literary analogy - metaphor theory tells us that these should therefore be rich in non-abstract analogies. The corpus is therefore made up of Brian Greene's 2004 *The Fabric of the Cosmos* and Marcus du Sautoy's 2003 *The Music of the Primes*, although we have subsequently tested the methodology on other popular science texts.

Through the procedure we describe in 3.1 below to analyse metaphor and analogy in these texts, we identify a range of domains which are unusually frequent in these texts and which are not pertinent to their subject matter (that is, not in the areas of physics, mathematics or general science). We then demonstrate in the remainder of section 3 that these domains are those analogies used systematically and consistently across the texts to elucidate and explicate the abstract concepts the books are focused on discussing. In order to do this, we identify all the semantic domains mentioned in these texts at very high levels of precision, using an annotation system built around the unprecedented detail found in the database of the *Historical Thesaurus*.

2. Semantic Annotation

Semantic tagging and annotation is, we argue, the best solution we have to address the problem of searching and aggregating large collections of textual data: at present, historians, literary scholars and other researchers must search texts and summarize their contents based on word forms. These forms are highly problematic, given that most of them in English refer to multiple senses – for example, the word

form "strike" has 181 *Historical Thesaurus* meaning entries in English, effectively inhibiting any large-scale automated research into the language of industrial action; "show" has 99 meanings, prohibiting effective searches on, say, theatrical metaphors or those of emotional displays. In such cases, much time and effort is expended in manually disambiguating and filtering search results and word statistics.

To resolve this problem, we use in this paper an early version of the Glasgow-Lancaster Semantic Annotation System, which we are currently developing at both of those universities. GL-SAS is a tool for annotating large corpora with meaning codes from the *Historical Thesaurus*, enabling us to search and aggregate data using the 236,000 precise meaning codes in that dataset, rather than imprecise word forms. These *Thesaurus* category codes are over one thousand times more precise than USAS, the current leader in semantic annotation in English corpus linguistics.⁸ The system automatically disambiguates these word meanings using existing computational disambiguation techniques alongside new context-dependent methods enabled by the *Historical Thesaurus'* dating codes and its fine-grained hierarchical structure. With our data showing that 60% of word forms in English refer to more than one meaning, and with some word forms referring to close to two hundred meanings, effective disambiguation is essential to GL-SAS.

3. Results

3.1. Methodology

The 600,000 word corpus we outline above were lemmatised and then processed through our annotation system, resulting in texts with each word being annotated with a *Historical Thesaurus* meaning code. We then aggregated those codes into a dataset which summarised the frequency of each meaning code in the text, and took that frequency list and compared it to a reference corpus made up of a 14m word corpus of random selections from Wikipedia, to provide a comparison against standard expository text. Our comparison was based on a log-likelihood significance measure,⁹ which identifies, to an acceptable degree, those semantic domains which are mentioned unusually frequently in our popular science texts by comparison to the reference corpus, and therefore indicates a text's "key" domains (where the log-likelihood values are greater than around 20)¹⁰ - those domains which reflect what a text is "about".¹¹

3.2. *The Fabric of the Cosmos*

Brian Greene's 2004 *The Fabric of the Cosmos* discusses theoretical physics and its relation to the concepts of space and time. Its key semantic domains are given in Table 1:

HT Category	Category Name	Log-Likelihood Value
01.05.07	Space	13655.8
01.05.07.01	Distance	6344.8
01.04.07.05.04.08	Photon	4912.5
01.05.06.07	Computation of time	3603.5
01.02.09.15	Spinning textiles	3193.5
03.11.03.01.08.02	Stringed instruments	2277.7
03.11.03.02.09.14	Pattern/design	1949.8
01.02.09.14.01.03	Woven fabric	1922.2

While the first four domains are within the *Thesaurus* categories which refer to the text's topic, and therefore expected, the next four (in bold) are not immediately relevant

to the book's topic. Looking for these domains in the text itself, chunked into 591 smaller files of 320 words each, we get a distribution like this:

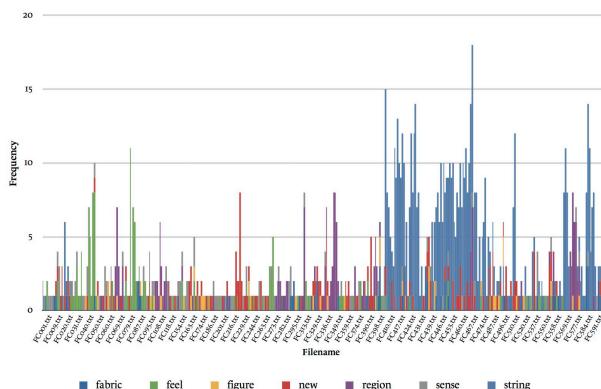


Fig. 1: Analogical textual clusters in *The Fabric of the Cosmos*, shown by frequency of key semantic domains

(Here, the Thesaurus codes have been replaced by words representing those categories, for ease of reading.)

The peak three-quarters of the way through the text indicates an area rich in mentions of textiles, and looking at this point in the text we find passages such as:

Since we speak of the ‘fabric’ of spacetime, the suggestion goes, maybe spacetime is stitched out of strings much as a shirt is stitched out of thread. That is, much as joining numerous threads together in an appropriate pattern produces a shirt’s fabric, maybe joining numerous strings together in an appropriate pattern produces what we commonly call spacetime’s fabric. Matter, like you and me, would then amount to additional agglomerations of vibrating strings.¹²

The areas we have identified through the log-likelihood analysis are therefore those areas rich in metaphors of fabric and strings (as other examples show) which are used by the author to discuss physics. We can therefore use this technique to pinpoint areas of significant use of metaphor or analogy in a text.

3.3. The Music of the Primes

As a check of the methodology, the same technique shows that in this particular book, which discusses prime number theory, there are highly key domains of *travel* and *landscape* in use alongside mathematical terms. Going to sections particularly rich in these domains gives analogical content over a long stretch, introduced by the following extract:

Gauss’s two-dimensional map of imaginary numbers charts the numbers that we shall feed into the zeta function. The north-south axis keeps track of how many steps we take in the imaginary direction, whilst the east west axis charts the real numbers. We can lay this map out flat on a table. What we want to do is to create a physical landscape situated in the space above this map. The shadow of the zeta function will then turn into a physical object whose peaks and valleys we can explore.¹³

4. Conclusion

We therefore demonstrate in this paper the use of a very fine-grained semantic annotation system, and establish the utility of such detailed annotations by describing a digital technique for discovering not only the existence of systematic metaphorical content but also its location and where it clusters. We believe that this result is significant in its own right, particularly for scholars of metaphor or cognitive linguistics, but we will also show that this represents only one of the uses to which highly-granular semantically annotated data can be put.

References

1. Kay, C., J. Roberts, M. Samuels, and I. Wotherspoon (eds). (2009). *Historical Thesaurus of the Oxford English Dictionary*. Oxford: Oxford University Press. See also historicalthesaurus.arts.gla.ac.uk.
2. Rayson, Paul. (2008). *From Key Words to Key Semantic Domains*. International Journal of Corpus Linguistics 13.4. 519-549.
3. Lakoff, George & Mark Johnson. (1980). *Metaphors We Live By*. Chicago: University of Chicago Press.
4. Gibbs, Raymond W. (2006a). *Introspection and Cognitive Linguistics: Should We Trust Our Own Intuitions?* Annual Review of Cognitive Linguistics 4(1). 135-151.
5. Evans, Vyvyan & Melanie Green. (2006). *Cognitive Linguistics: An Introduction*. Edinburgh: Edinburgh University Press. Page 780.
6. Alexander, Marc & Christian Kay. (2011) [2010]. *Mapping Metaphors Across Time with the Historical Thesaurus*. Conference paper at Helsinki Corpus Festival: The Past, Present, and Future of English Historical Corpora, University of Helsinki, Finland. Based on an earlier paper at The 3rd UK Cognitive Linguistics Conference, University of Hertfordshire.
7. Alexander, Marc. (2011). *Meaning Construction in Popular Science An Investigation into Cognitive, Digital, and Empirical Approaches to Discourse Reification*. University of Glasgow: Ph.D. thesis.
8. ucrel.lancs.ac.uk/usas
9. Dunning, Ted. (1993). *Accurate methods for the statistics of surprise and coincidence*. Computational Linguistics 19(1). 61-74.
10. Rayson, Paul, Damon Berridge, & Brian Francis. (2004). *Extending the Cochran Rule for the Comparison of Word Frequencies between Corpora*. 7th International Conference on Statistical Analysis of Textual Data.
11. McIntyre, Dan & Brian Walker. (2010). *How can Corpora be Used to Explore the Language of Poetry and Drama?* In Anne O’Keeffe & Michael McCarthy (eds.), *The Routledge Handbook of Corpus Linguistics*. London: Routledge. 516-530.
12. Greene, Brian R. (2004). *The Fabric of the Cosmos: Space, Time and the Texture of Reality*. Alfred A Knopf: New York. Page 486-7.
13. du Sautoy, Marcus. (2003). *The Music of the Primes: Why an Unsolved Problem in Mathematics Matters*. London: Harper Perennial. Page 85.

The Cryptic Novel: A Computational Taxonomy of the Eighteenth-Century Literary Field

Algee-Hewitt, Mark

mark.algee-hewitt@stanford.edu
Stanford University

Eidem, Laura

lmeidem@stanford.edu
Stanford University

Heuser, Ryan

heuser@stanford.edu
Stanford University

Law, Anita

anital@stanford.edu
Stanford University

Llewellyn, Tanya

tanya.llewellyn@gmail.com
Stanford University

Overview

This project employs machine learning and other text-mining based clustering techniques to study the relationship between taxonomic systems for categorizing prose fiction in the eighteenth century. Our goal is to challenge narratives of the so-called “rise” of the novel (Watt, 1957¹; McKeon, 1987²; Richetti 1999³) that trace a continuity between the prose fiction that emerged in the eighteenth-century and the novel as a fully realized critical object in the late nineteenth century. Our project instead uses statistical modeling methods to theorize an alternate history for the novel: not, as proposed in Franco Moretti’s *Graphs, Maps, Trees*⁴, as a history of genres, but instead as a history of genre labels—that is, a history formed by the interaction between paratextual, self-identified literary labels that circulated in the eighteenth-century marketplace and the texts that they defined. Our project, which is currently in its final stages at Stanford University’s Literary Lab, bridges the historical and computational to enact a form of literary history that puts pressure on traditional theories focused on charting the rise of the novel. The project is deeply invested in using digital and statistical methods to excavate the nuances in the ways in which eighteenth-century genres evolved. Instead of analyzing texts via contemporary genre categorizations, we recover the relationship between the self-applied taxonomy of eighteenth-century prose fiction and the lexical and semantic features of the texts themselves to recuperate a historically-determined literary field that has been largely overlooked (or noted and elided) by modern criticism. Drawing from a corpus of 2,385 digitized texts from the Eighteenth Century Collections Online (ECCO) database, we ask how new computational methods can aid us in uncovering what kinds of work self-identified genre labels in titles (specifically “story,” “history,” “tale,” “letters,” “romance,” “life,” “adventure,” and “novel”) accomplished within the literary marketplace of the long eighteenth-century, and how that kind of work is distinct from critically- or retroactively-designated genres (“gothic,” “jacobin,” “epistolary,” “historical,” “it-narrative,” “oriental tale.”)

Aims

Our project seeks to answer a series of critical questions about the taxonomic systems of eighteenth-century fiction using computer generated models. Do generic labels formally differentiate separate kinds of writing in a useful way (such that a “novel” is lexically, semantically or formally different from a “romance”) or do they merely function as signals within the marketplace itself, so that there is no textual or formal differences between a “tale” or a “story” beyond mere marketplace convention? How does the relationship between different genre labels change over time, and how can it assist us in understanding the evolution of titles as a representative labeling system? This approach takes advantage of recent advances in probabilistic modeling to recover the meaning of labels that have been homogeneously condensed under the rubric of novel; it simplifies the complexity of the word “novel,” with all its attendant genres and subgenres, to make the field that “novel” inhabits more complex. Developing from these research questions, our method-driven aims are: (1) to detect and assess large-scale trends in the development of and relation between genre labels across time (2) to isolate formal differences in the corresponding full texts identified under these categorical labels, and (3) to compare these differences to a corpus of texts categorized according to modern genre designations.

Methods and Narrative

In the early stages of our project, we began with an exploratory approach to our data, using statistical, unsupervised learning techniques to identify word patterns and clusters within the texts, prior to classifying the texts’ content onto the label categories. Our initial approach sought to determine if some underlying structure existed among these texts that we could then map onto a taxonomic system, such as genre, or in our case, title labels. Using a series of clustering techniques built off of different feature sets (most frequent words, most

distinctive words, etc), we found that the assumptions that provided the foundation for our project were vindicated: not only did the titles under each label cluster together based on the Most Frequent Words in each text, but there were intriguing differences between clusters. For instance, though a coherent cluster of “history” is present from the first quarter of the century, definitively clustered labels “novel” and “romance” don’t emerge until later in the century. These differences opened our research to a new set of concerns, such as the relative stability of each category, or the ways in which the labels individually and collectively change across time. Indeed, a major finding of this phase was the dramatic influence time itself had over every other variable we considered. Such a finding strengthened the presence of a literary-historical component of our research and prompted us to pay closer attention to the divisions of time we were employing.

In our current work, we have employed supervised learning techniques and machine classification to specifically interrogate the relationship between text and label: which labels have a high level of cohesion, and thus predictability, and which labels are less cohesive, and were perhaps more tolerant of formal deviation and experimentation? To answer this question, we employed a discriminant function analysis to determine if texts could be reliably algorithmically classified into their labels, treated in our analysis as a taxonomic system. These results were validated through a leave-one-out cross validation to determine the strength of our predictive model. . The results confirmed the findings of the PCA: our model performed better than chance, categorizing each of our labels, on average, 75% correctly. Another dimension of our study was to investigate differences on a label-by-label basis: what does it mean, textually, that a “life” is harder to algorithmically classify than a “romance”? Did difficulties in categorization have anything to do with the textual heterogeneity within the labels themselves? These questions prompted us toward another set of analyses designed to evaluate the self-similarity of each label. We divided each text in our corpus into ten equal parts, measuring the distance from each part to the other in a matrix that, when averaged, resulted in a distance score for each. The findings from this activity helped to both clarify and complicate the story that was emerging regarding labels over the course of the century.

Our most recent work has taken us in a more broadly comparative direction: to examine the textual and larger structural differences between our current corpus of texts, classified according to generic labels, and another corpus we have compiled from modern, genre-focused bibliographies. Using the same techniques of machine learning and self-similarity, we have started to interrogate the logics by which generic labels organize and describe the literary field of the eighteenth century in ways we have yet to discover.

Results

Our initial results proved the value and worth of studying this type of paratextual literary categorization. The temporalization of data allowed by our method lets us observe, first, the simultaneous dominance and differentiation of “novel” and “tale” at the end of the century, leading us to speculate, as Anthony Jarrells (2012)⁵ has, on twin “rises” of genre in this century as opposed to one rise; that is, the eighteenth-century appears to be as much a story of the rise of “novel” as it is of the rise of “tale.” The result of our machine learning classifications and self-similarity form a compelling argument for the ways in which, throughout the eighteenth-century, the field of prose fiction underwent a transformation through the processes of both differentiation and consolidation. From these results, we argue that the novel can be viewed as merely one branch, rather than the single formal end to which all eighteenth-century prose fiction trends. Other important results, observed from our PCA charts, indicate a general movement across all generically-labelled texts from the usage of words that indicate exteriority (such as “law,” “city,” “army,” “money,” “order,” and “public”) to words that indicate interiority (such as “lover,” “marry,” “imagine,” “woman,” “write” and “read”). In the final stages of the project, we are working to apply similar computational

tools to genres conventionalized by modern criticism -- such as the Gothic novel or Oriental tale -- to examine how such categorizations relate to self-applied genre labels. Preliminary results indicate that, after sampling our texts and controlling for time period and author, generic labels classify with a success rate comparable to texts designated as such by critics. The question then becomes not whether genre labels organize more successfully than conventional genres, but whether and how they organize differently--if generic labels can tell us something about the trajectory and organization of formal literary history that genre cannot. Our presentation will include a more detailed discussion of our methods and results and of the implications of the latter for traditional teleological narratives of eighteenth-century prose fiction

References

1. Watt, Ian (1957). *The Rise of the Novel: Studies in Defoe, Richardson, Fielding*. London: Chatto and Windus.
2. McKeon, Michael (1987). *The Origins of the English Novel: 1600-1740*. Baltimore: Johns Hopkins UP
3. Richetti, John (1999). *The English Novel in History, 1700-1780*. London: Routledge.
4. Moretti, Franco (2004). *Graphs, Maps, Trees: Abstract Models for Literary History*. New York: Verso
5. Jarrells, Anthony (2012). *After Novels: Short Fictional Forms and the Rise of the Tale*. The Oxford History of the Novel in English. 12 vols. Vol. 2. Eds. Peter Garside and Karen O'Brien. Oxford: Oxford UP.

The Stanford Literary Lab Tranhistorical Poetry Project Phase II: Metrical Form

Algee-Hewitt, Mark

Stanford University, United States of America

Heuser, Ryan

Stanford University, United States of America

Kraxenberger, Maria

Stanford University, United States of America

Porter, J.D.

Stanford University, United States of America

Sensenbaugh, Jonny

Stanford University, United States of America

Tackett, Justin

Stanford University, United States of America

To date, most studies that foreground quantitative analyses of literature have focused exclusively on prose writing (the novel in particular) rather than poetry (Stanford Literary Lab 2011, Clement 2008). In part, this state of affairs is due to poetry's highly figurative language and complex communicative intent, which poses acute problems for text mining and similar quantitative analyses that rely on lexical and semantic meaning and whose methodological origins lie in the "hard" sciences (Pasanek and Sculley 2008, Bei 2008). Poetry, however, offers a unique subject for quantitative analysis, independent of lexical and semantic meaning, that is largely absent from prose works: meter. The practice of scansion is an ancient study that precedes the advent of the English language. It has, nevertheless, always consisted mainly of counting, sorting, and indexing words and word components, endeavors to which quantitative analysis is especially attuned. Despite this apparent sympathy between metrics and quantitative analysis, however, the algorithmic detection of the complexities of meter remains outside of the current capabilities of the Digital Humanities. The irregularity of stress, syllabic schemes, and the rule-bending nature of poetic diction runs counter to the binary presence/absence process of most computational analysis.

In the second phase of the Stanford Literary Lab's multi-year ongoing project to create a system for detecting the formal features of poetry, we have focused our attention on the question of meter. Using a new method that combines a series of rule-based analyses with an iterative probabilistic-based classification algorithm, we can now detect, with a high degree of accuracy, both the meter and line length of individual poems. We have trained our algorithm to recognize individual metrical feet, such as iambic, dactyls, anapests, trochees, and spondees, and to combine these identifications with a signal-processing approach to the entire poem to classify the overall metrical scheme of any given poem. We have trained and tested our algorithm on a corpus of over 300,000 English language poems from the late medieval period to the twentieth century. Moreover, we have also applied our algorithm to multilingual corpora: in our presentation we will demonstrate how our methods are effective on German and French, as well as English poetic forms.

This project builds upon the first phase of our project, presented to the Alliance of Digital Humanities Organizations conference in Lincoln, Nebraska, which successfully sought to recognize the syllabic scheme patterns in poetic lines. The overarching goal of our project as a whole is perhaps simple to state, but challenging to execute: we seek to create a program that can automatically identify poetic forms. In other words, we are in the process of designing a program that can read any number of poems and tell us their exact syllable count, meter, rhyme scheme, use of traditional forms (e.g., sonnet, ballad, sestina), etc. Success in creating such a program would represent an important tool for scholars in the Digital Humanities, and would offer the ability to:

- 1) Create an inventory of all poetic forms, traditional and untraditional.
- 2) Trace the history of poetic forms, including:
 - variation *among* poetic forms (e.g., Which forms were most popular in a given period? Were some periods more formally diverse than others? How does form diversity change over time?)
 - variation *within* poetic forms (e.g., How do the forms themselves change over time? Are sonnets more metrically rigorous in one era than another?)
- 3) Better understand the relationship between form and meaning by relating analyses of scansion with those of lexical and semantic meaning.
- 4) Provide distant readings to help generate and/or support new close readings.

In constructing such a program, we break down the task of recognizing metrical schemes into the simpler, but by no means simple, components of recognizing: 1) Syllable Scheme; 2) Beat Scheme; 3) Rhyme Scheme; 4) Metrical Scheme; and then 5) matching any combination of these categories to a tradition name (e.g., sonnet, heroic couplets, rhyme royale, etc.) if one exists to describe it.

We began by designing a program that could accurately detect the number of syllables in a given line of poetry (item 1 above) because we believed it would be the most straightforward element to analyze. We additionally realized that if we limited our sample according to metrical foot, then syllable count would present a shortcut toward detecting a rough approximation of meter. For example, if we started by analyzing only iambic poems (as recognized by human readers), and our syllabifier counted 9-11 syllables in each line of a given poem, then we could have reasonable assurance that the poem was written in pentameter. A similar process could be applied to other metrical patterns. In training our syllabifier, we purposely limited our sample to poems composed from roughly the late sixteenth century to the late nineteenth century because metrical forms were most stable and recognizable in this period.

Building upon this earlier success, we have combined our ability to recognize syllabic scheme with a complex approach to meter that has, at present, an over 80% success rate in correctly classifying meter. Our presentation will discuss how our algorithm was built, the specific challenges that we faced (e.g., elision, extrametrical syllables, feminine endings, foreign words, and other features of meter that are commonly acceptable in the practice of poetics but that our program had

difficulty overcoming), and we will present the results of our application of this technique to a large, multi-lingual corpus that shows the historical shape of various metrical forms important to European poetry from the sixteenth to the late nineteenth century. Our initial analysis, using only syllabic scheme, revealed a number of significant and unexpected observations, including the fact that the use of pentameter peaked around the middle of the seventeenth century and has been on the decline ever since, as well as the fact that the use of tetrameter has been reciprocally on the rise since the early eighteenth century and is today equally as popular as pentameter. In this second phase, we expand the results of this analysis to show the historical prominence of the sonnet and heroic couplet forms, the transnational inheritance of metrical form and the history of iambic pentameter in English poetry.

We believe that what we have achieved with this work so far will aid in future quantitative and digital work on poetry, a lacuna that represents a critical problem for the use of digital humanities in the study of literature, given the enormous significance of the poetic tradition within literary studies. We also believe, moreover, that our program has application well beyond the study of poetry and could help to analyze and detect metrical schemes in song, drama, and prose as well, generating topics of analysis that likely would remain undetected without quantitative analysis.

Common Container Correlation: A Simple Method for the Extraction of Structural Models from Statistical Data

Alvarado, Rafael

ontoligent@gmail.com

University of Virginia

The use of topological graphs, or networks, to represent and analyze the semantic contents of source materials, such as texts and images, has become a signature contribution by the digital humanities to the humanities in general. Specific techniques, such as topic modeling and network analysis, and general approaches, such as macroanalysis and distant reading, exemplify the popularity and effectiveness of methods based on the graph theoretical representation and statistical modeling of cultural materials. However, because of their mathematical complexity and their focus on very large corpora of texts, these methods are beyond the reach of many humanists interested in the interpretation of smaller sets of source materials for cultural meaning. They are also suspect since they introduce ontological commitments that both elide traditional notions of human agency and reframe culture as a set of abstract, metrical dimensions. In this talk, I introduce “common container correlation” (C3) as a relatively simple and transparent interpretive method for the graph theoretical analysis of source materials that may be practiced by both students and more advanced researchers to excavate and make sense of cultural models implicit in textual materials.

C3 may be described as a variation of co-occurrence analysis designed to take advantage of the abundance of encoded cultural materials available to the digital humanist and to allow for the analysis of small sample sizes, such as individual texts. Formally, a common container correlation is just a link, or edge, that is asserted between any two items, regarded as nodes or vertices, that are contained within the same structural container. The set of all such links produces a graph of nodes and links based on their co-occurrence in a common container. In some cases these graphs will have meaning—that is, they will exhibit patterns that lend themselves to structuralist and other forms of interpretive analysis. These patterns may sometimes be correlated with psychological, sociological, or material causes that will be of interest to the humanist.

For example, in a novel marked up with TEI-based schema, we may choose to define the paragraphs of the text as container elements and tagged references to proper names as contained elements. We then assert that all named agents in a given paragraph are related to each other (in the special sense of co-occurring). The set of all of these assertions for all paragraphs will produce a kind of social graph that may then be visualized and analyzed in structural terms. In such a case, it may emerge that two characters consistently appear on opposite sides in multiple instances of a force-directed representation of the graph. This may be evidence of a structural opposition that will have emerged from the statistical distribution of the selected elements. Other approaches may use other container elements, such as scenes, and combinations of contained elements, such as places and people.

The C3 method is easy to implement using available tools. Container and contained elements in XML encoded materials may be extracted using simple XPath statements (by means of a variety of tools) and dumped into tables with columns for container IDs and contained IDs. Such tables may then be transformed using simple SQL queries into various graph data formats for visualization and analysis in tools such as Gephi, GraphViz, SHIVA, and D3. Depending on the intention of the user, the resultant graph may or may not reflect the frequency of edges and vertices in the source data.

In this talk I will describe the C3 method using examples taken from three digital humanities projects with which I have been associated. First, I will describe the application of the method to rhetorical figures (containers) and characters (contained elements) using data from the *Princeton Charette Project*. Second, I will describe how undergraduates in an introductory digital humanities course at the University of Virginia created a database relating characters and paragraphs in Austen’s *Persuasion*. Third, I will describe the use of the method in Stephen Railton’s *Digital Yoknapatawpha Project*, drawing on data correlating scene containers to people and places as contained elements in William Faulkner’s corpus of fiction. In each case, I will explore the interpretive implications of the algorithms used to visualize the data, taking particular care to describe the specific steps involved in going from markup to data representation to visualization to interpretation. In this way, I hope to connect the discourse on data-driven textual analysis to traditional interpretive methods, such as close reading and structural analysis, in order to produce a genuinely humanistic use of quantitative methods that does not alienate the researcher from the tools of interpretation.

Rethinking Recommendations: Digital Tools for Art Discovery

Andrew, Liam

landrew@mit.edu

Massachusetts Institute of Technology, United States of America

Gonzalez, Desi

designoz@mit.edu

Massachusetts Institute of Technology, United States of America

Automatic discovery and recommendation systems are often designed with one of two audience groups in mind: in academia, the target is the dedicated researcher who actively seeks out particular sources, whereas companies like Amazon or Netflix design recommendations for the casual, passive browser, with convenience as the top priority. Often, however, a user is both browser and researcher in separate tabs; while diving into research in a scholarly database, a user can simultaneously peruse news aggregators or Amazon. For-profit companies often recommend cultural products such as books and movies, but do so with a single goal—increasing the company’s profit. As digital humanists, we should rethink the structure of recommendation algorithms to make them more

appropriate for audiences interested in deeper explorations of cultural heritage.

At HyperStudio, we are investigating how digital tools can encourage discovery and serendipity in the humanities, with a focus on art objects and museum collections. For this short paper session, we propose to share our research on the process of discovery, assessing algorithms used in research and recommendation tools on both scholarship and industry platforms. We will survey existing projects that allow scholars and casual users alike to discover new art. We will also discuss a tool that we are building, tentatively titled ArtX, that empowers users to discover cultural events, exhibitions, and art objects in the Boston area. Informed by our theoretical research into cultural recommendation systems, we are prototyping and testing this tool this spring and will be sharing our results at DH2014.

Recommendation systems are typically divided into two approaches: collaborative filtering and content-based filtering.¹ While many digital tools use these in combination, here we outline the approaches and their limitations separately. Content-based filtering approaches, such as traditional tagging systems, look at the properties of the content rather than the user. Whether human- or machine-powered, tagging involves inferring what an object is “about” and how one might search for it, and assigning keywords of names, topics, or entities. The act of classifying culture is by its nature restrictive; when an art object is called “surrealist” or “American,” it is placed in a particular discourse and others are implicitly excluded. Even outside-the-box descriptions such as “hazy” are just different boxes. Artsy’s “Art Genome Project” offers a more nuanced approach to tagging (with gradients from 0 to 100, rather than 0 to 1), but this runs into the same problem.² When an authoritative institution such as a museum produces tags, the tagging system lacks dynamism. User-generated tagging, or folksonomies, add a dynamic element but require that users actively and continually contribute to building up the tags, a process that is difficult to maintain.

Collaborative filtering attempts to sidestep these limitations, focusing instead on the user and their online behavior, similar users, and social networks. User history-based approaches like Amazon’s maximize efficiency at the sake of variety, assuming that a user has no desire to try something new. Social curation tools such as Curiator, ArtStack, Pinterest and Tumblr allow users to build their own collections and share with others, but they perpetuate what is already popular or the most reblogged. Collaborative filtering may work when shopping for a product, but risks creating a filter bubble for art. It shepherds audiences into identical routes of understanding, stifling productive conversation and undiscovered treasures in the process. At the heart of these approaches is the notion that more personalization leads to higher quality, and that existing networks and canons should be reinforced; these are meaningful signals, but they should not be the only ones.

One alternate approach is to include a serendipitous chance in the discovery process. The role of serendipity in scholarly research has been a growing topic of investigation in recent years.³ Serendipity has historically played a significant role in science, mathematics, and the humanities. As resources are increasingly digitized, an oft-cited lament is the lack of serendipity, yearning for the days when a scholar would go to the library stacks looking for one book and happen upon another that sparks his or her thinking in new directions.

While serendipity is chance-based and cannot be controlled, perhaps it can be engineered. A few existing digital humanities and cultural heritage projects experiment with engineering serendipity. Serendip-o-matic, launched in August 2013, aims to re-incorporate chance into the scholarly research process. On the website, users input a text; the tool identifies key words in the text and responds with primary source images from several online collections. The goal of Serendip-o-matic is to yield happy accidents for a wide range of users, whether students in search of inspiration for a paper topic or scholars looking for materials to enliven a current project.⁴ Another example is Magic Tate Ball, a mobile application designed by digital studio Thought Den to encourage a general audience to discover works of art in the Tate’s collection. Using GPS location, time

of day, weather, and analysis of ambient noise, the application returns an artwork, explaining why this work was selected and providing content that allows the user to learn more.⁵ Magic Tate Ball enables users to engage with works they would not have sought out otherwise while infusing play in the discovery process.

At HyperStudio, we hope to incorporate a similar sense of serendipity in ArtX. Serendipity has the dual advantage of skirting traditional boundaries and adding a playful element to the user experience, which serves both browser and researcher. As we aim to make meaningful and creative connections between the art objects that comprise our past and the events of the present, we believe we can incorporate both audience groups without sacrificing archival rigor. To do so, we will need a holistic, audience-centered approach to digital curation and recommendation.

To achieve this goal, we plan to start small. Through specific partnerships with museums in Boston, we are building a closed and controlled system that can serve as a testing ground for new models of recommendation. Free from industry demands such as growth and scale, we can perfect our schemas and our assumptions before expanding to other institutions. We are also hopeful about creating a collaborative, open-source approach to art recommendation, particularly given the close secrecy with which proprietary recommendation algorithms are guarded. By encouraging open conversation around the ways we recommend art, we may find unique approaches and ways in which current recommendation systems are insufficient or misleading.

We have many questions and challenges ahead. It will be important to understand our audience: How much control over the discovery process do users want, and how can we best balance the sliding scale between browser and researcher? We expect our primary audience to be Boston-area residents and university communities—a casual but informed audience that bridges aspects of both. We hope to instill a scholar’s depth of interest and rigor in the casual user and we hope scholars too can employ the tool as serendipitous inspiration for their own work. But how transparent can we be about the logic behind our recommendations? How can we scale such a strategy, connecting artworks to books, lectures, music, movements and ideas?

Perhaps most importantly, while we have explained “why serendipity,” we must address the “how.” Serendipity involves more than simply selecting objects at random, but what signals are important? How can we prime a user for the mindset of serendipitous discovery, rather than rote research? Moreover, is it truly serendipitous if we are closely engineering the suggestion? We look forward to addressing these questions, but with care to not create our own faulty algorithms. One of the challenges in this process is to avoid reducing cultural objects to the level of products, and museum audiences to consumers. Looking past the current limitations of discovery will be vital for generating new connections and ideas.

References

1. A.A. Kardan and M. Ebrahimi, *A novel approach to hybrid recommendation systems*, Information
2. Interview: Matthew Israel on *The Art Genome Project*, September 21, 2013, Museum Geek, museumgeek.wordpress.com/2012/09/21/interview-matthew-israel-on-the-art-genome-project.
3. Scholarship includes Allen Edward Foster and Nigel Ford, “Serendipity and Information Seeking: An Empirical Study,” *Journal of Documentation*, 59 (2003): 3, pp. 321-340; Sebastian Chan, “Tagging and Searching – Serendipity and museum collection databases” (paper presented at the annual meeting for Museums and the Web, San Francisco, California, April 11-14, 2007); and Anabel Quan-Haase and Kim Martin, “Digital Humanities: The Continuing Role of Serendipity in Historical Research” (paper presented at the annual meeting for iConference, Toronto, Canada, February 7-10, 2012).
4. One Week | One Tool Team Launches Serendip-o-matic, Roy Rosenzweig Center for History and New Media, Friday,

August 2, 2013, chnm.gmu.edu/news/one-week-one-tool-team-launches-serendip-o-matic.

5. **Ben Templeton** (2012), *Mobile Culture and the Magic Tate Ball*, The Guardian, July 16, www.theguardian.com/culture-professionals-network/culture-professionals-blog/2012/jul/16/mobile-culture-magic-tate-ball-app.

L'édition numérique – système d'organisation des connaissances avec les outils sémantiques

Andréys, Clémence

Université de Franche-Comté, France

Borel, Clément

Université de Franche-Comté

Roxin, Ioan

Université de Franche-Comté

La culture numérique, ses pratiques et ses exigences ont profondément modifié les pratiques de lecture en introduisant de nouvelles manières de lire. Cela a suscité des changements dans l'édition qui a développé de nouveaux supports et de nouveaux formats à l'aide de technologies capables de transformer l'environnement numérique en environnement de lecture. Avec les outils sémantiques émerge un nouveau modèle d'édition numérique qui est bien plus qu'une simple base de connaissances.

Le projet Descartes est un projet d'édition numérique mené par l'équipe Objets et Usages Numériques (Laboratoire ELLIADD) et le Centre de Documentation et Bibliographie Philosophiques de l'Université de Franche-Comté au début des années 2000. Il va bien au-delà de la plate-forme de publication, fonctionnalité largement répandue : il s'est interrogé sur les apports des technologies numériques et a élaboré un dispositif ayant pour mission de compléter, voire de renouveler l'expérience de lecture. Ce projet a trouvé un prolongement dans une thèse de doctorat intitulée « Outils sémantiques au service du livre numérique, modélisation et visualisation des liens transtextuels ». L'objectif était de montrer de quelle façon les technologies du web sémantique pouvaient être utilisées pour le développement d'une plate-forme de lecture numérique en tissant des liens entre des textes. La structuration des connaissances se fait selon trois niveaux : informationnel, descriptif et sémantique. Le niveau informationnel est composé de l'ensemble des ressources numériques primaires (e.g. ressources textuelles et iconographiques). Les métadonnées sur ces documents primaires se retrouvent au deuxième niveau, descriptif, et sont enregistrées en triplets RDF comme une base de connaissances assertionnelles. Le troisième niveau, sémantique, contient les connaissances théoriques du corpus formalisées en ontologies (RDFS, OWL-DL). Cette réflexion a nécessité la modélisation d'une ontologie de la transtextualité, appelée TROW (Web Ontology of TRanstextuality), qui définit les liaisons qui unissent les différents textes à l'intérieur d'un corpus. Ces liaisons, enregistrées sous la forme de métadonnées, permettent de naviguer efficacement dans le corpus et offrent une approche différente grâce à une nouvelle organisation et présentation des informations. La machine complète le travail de l'utilisateur dans la mesure où elle interprète les relations qui existent entre les textes et peut ainsi inférer des liaisons inédites. Cela a conduit à la création d'une interface de lecture adaptée à la navigation transtextuelle, en particulier via une disposition parallèle des textes et une représentation des liens profonds sous forme de graphe. L'interface est capable d'afficher un graphe des liaisons transtextuelles à l'activation d'un lien et permet la lecture simultanée d'un texte de référence et d'un texte cible.

Le travail que nous aimeraions présenter constitue la prochaine étape de cette réflexion sur la transmission du savoir grâce aux technologies du web sémantique. Il s'inscrit

dans une collaboration entre chercheurs qui ouvre le projet Descartes, ancré en philosophie, vers d'autres disciplines pour aboutir à un projet interdisciplinaire où interviennent sciences de l'information et de la communication, informatique, histoire, études chinoises et études germaniques. Il vise le passage d'un outil pour la lecture de textes à un outil permettant de voyager dans une base de données qui ne sera pas seulement composée de textes. Il s'agit en effet d'éditer un corpus d'archives composé entre autres de cartes, de photographies et de textes sur la présence allemande en Chine au tournant du vingtième siècle. Nous nous intéresserons en premier lieu à un ensemble de cartes postales qui véhiculent des représentations de la Chine et de ses habitants dans l'Allemagne wilhelminienne. Nous assurerons la description de ces sources primaires dans la mesure où nous définirons, structurerons et associerons les métadonnées correspondant aux cartes postales afin d'établir une nouvelle ontologie.

L'objectif est similaire à celui du projet Descartes : accompagner l'utilisateur dans la découverte d'un ensemble de documents, l'aider dans son analyse en tissant des liens entre les ressources documentaires grâce aux possibilités offertes par les technologies du web sémantique. Cela implique d'utiliser l'ontologie TROW comme une base, tout en la complétant, voire en la modifiant pour l'adapter à tous les genres de documents du corpus et de créer plusieurs ontologies, afin d'assurer l'interopérabilité des liaisons.

L'infrastructure doit permettre une visualisation interactive dans la mesure où les représentations visuelles sont considérées comme de véritables supports de la pensée. Elle doit être capable de gérer le déploiement d'archives numériques qui ne prend pas la forme d'une narration hypertextuelle, mais celle d'une représentation cartographique. L'utilisateur reste libre dans sa démarche de découverte et d'appropriation des connaissances : il n'est pas question d'imposer un sens de lecture et une interprétation des documents. La machine propose des connexions nouvelles, mais c'est à l'utilisateur de jouer un rôle actif dans l'acquisition et le partage des connaissances. On passe d'un espace de stockage de documents à un espace interactif qui associe les fonctions du web sémantique et la participation des utilisateurs. Chaque utilisateur compose sa propre sélection et crée son propre parcours dans la base de données. Il a la possibilité d'évaluer la pertinence des sources, d'annoter, de commenter, d'échanger. Ainsi chaque utilisateur devient l'auteur d'une mémoire personnalisée et prend part simultanément à l'écriture de la mémoire collective. Il faut également poursuivre les recherches qui concernent l'ergonomie de l'interface afin de répondre à l'élargissement des fonctionnalités de l'outil et aux besoins de l'utilisateur.

Il est aussi important de souligner que cet outil vise à développer une passerelle entre le monde universitaire et le grand public en diffusant la recherche en sciences humaines et sociales par un média numérique. Il a pour but d'assurer une diffusion et une réception plus larges d'archives numérisées. Il doit garantir une meilleure visibilité des travaux de recherche au sein de la communauté universitaire, mais aussi une vulgarisation des connaissances scientifiques, une démocratisation du savoir en rendant l'utilisateur actif et en favorisant le travail collaboratif. Il s'agit de penser la transmission du savoir au-delà d'une simple numérisation des sources et d'envisager l'apport des technologies du web sémantique dans la constitution d'une mémoire individuelle et collective.

References

Berners-Lee, Tim, Hendler, James et Lassila, Ora (2001), The Semantic Web, A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities, Scientific American, mai 2001, p. 34-43.

Borel, Clément, Royin, Ioan (2013), Ontologie de l'intertextualité. Le cas des corpus numériques philosophiques, in Pratiques et usages numériques. H2PTM'13, Lavoisier, Paris, p. 281-290.

Kaplan, Frédéric, (2013) L'origine médiévale de l'hyperlien, des pointeurs et des smileys avril 19 2013. Disponible à

l'adresse : fkaplan.wordpress.com/2013/04/19/lorigine-medievale-de-lhyperlien-des-pointeurs-et-des-smileys

Ong, J. Walter (2002), *Orality and Literacy. The Technologizing of the World*, Routledge, New York.

Pédaueque, Roger T. (2006), *Le document à la lumière du numérique*, Caen, C&F éditions.

Peirce, Charles Sanders (1978), *Écrits sur le signe*, Seuil, Paris.

Tricot, Christophe (2006), *Cartographie Sémantique, des connaissances à la carte*, thèse de doctorat, Université de Savoie.

From the Archimedes' Palimpsest to the Vercelli Book: Dual Correlation Pattern Recognition and Probabilistic Network Approaches to Paleography in Damaged Manuscripts

Anthony, Eleanor Chamberlain

University of Mississippi, United States of America

The human capacity to visually discern and detect information is remarkably powerful. This ability forms the basis for the field of paleography, specifically in terms of text identification. However, while traditional paleographical methods yield impressive results for data characterization, a mathematical approach to identification is desirable as a means of additional verification and as a basis for comparison between varying proposals of text identities, especially in the case of damaged manuscripts in which text is wholly or partially illegible. Executed judiciously, mathematical approaches can aid the transcription efforts for manuscripts and encourage discussion of text identity with appeal to statistical distribution, based on both visual and contextual information. !

Derek Walvoord initially employed such an approach for the Archimedes' Palimpsest in 2004.1 He developed a correlation recognition program and probabilistic network system for data analysis that utilized both visual and contextual information in order to create character identity distributions for the data of interest. I am currently working to adapt and improve his approach in order to analyze image data for the Vercelli Book. The oldest existing record of Anglo-Saxon poetry and homilies, this 10th-century manuscript was imaged in March 2013 by a team from the Lazarus Project at the Archivio Capitolare in Vercelli. Led by Dr. Gregory Heyworth, the team was composed of Dr. Roger Easton, Kenneth Boydston and Dr. Keith Knox, accompanied by several students from the University of Mississippi, including myself. Utilizing reflective and transmissive LED arrays in 12 wavebands ranging over the near-ultraviolet, visible, and near-infrared regions, we imaged the manuscript with the aim of penetrating the chemical reagent applied in the 19th century that obscures sections of the text. !

In order to determine the identities of characters and character fragments which image rendering do not conclusively reveal, I am designing a neural network system which will make use of correlation filtering to classify the data, similar to Walvoord's system; however, by utilizing subsampling of the initial and subsequent output layers, I will be able to meaningfully analyze fine details of the data in a way that Walvoord's system did not allow. Moreover, my updated system will be able to weight the importance of particular features in determining class membership via a learning architecture, such that each training image improves the classifier function. This method should overcome problems traditionally associated with correlation filtering methods such as blur, lighting differences, and lack of contrast because of its subsampling and weighting mechanisms. Depending upon the results of this neural network approach of classification for the Vercelli Book data, I will evaluate whether employing a probabilistic network

system similar to that used for the Archimedes' Palimpsest will significantly improve overall data classification and proceed with research accordingly.!

This system would serve to bridge the gap between our knowledge of the Vercelli Book's contents and the information available to us, providing a probabilistically accurate transcription of data associated with a piece of the humanities corpus that has been obscured from view for over a century. This invaluable data will contribute place and context to our current conversations considering Anglo Saxon literature, language, and culture, as well as manuscripts in general. Moreover, this system will ultimately serve to empower paleographers and computer scientists alike by exploring the relationship between the visual acuity and classification knowledge associated with traditional transcription efforts and neural network-based approaches for mathematical ones.

References

Walvoord, Derek J. (2008). "Advanced Correlation-Based Character Recognition Applied to ! the Archimedes Palimpsest." PhD diss., Rochester Institute of Technology.

Sermanet, Pierre et al. (2012) "Convolutional Neural Networks Applied to House Number Digit Classification." Paper presented at the 21st International Conference on Pattern Recognition, Tsukuba, Japan, November 11-15.

Tracing Workflow of a Digital Scholar

Antonijevic, Smiljana

Roskilde University, Denmark

Stern Cahoy, Ellysa

Penn State University, USA

This paper presents findings of an Andrew W. Mellon Foundation-funded project conducted at Penn State University in the period April 2012-June 2013. The project explored scholarly workflow of the Penn State faculty across the sciences, humanities, and social sciences, focusing on the integration of digital technologies at all stages of a research lifecycle—from collecting and analyzing data, over managing and storing research materials, to writing up and sharing research findings. The study also examined scholars' attitudes towards the use digital technologies in their research practice, as well as the level of institutional support available to them in developing and implementing digital research skills. This paper harvests a comparative multidisciplinary perspective of our study in order to explore specificities of humanities scholars' digital workflow, providing a ground to identify and develop a software and service architecture that supports those practices. Therefore, while focusing on current findings, the paper briefly highlights the future trajectory of our study, as well as planned next steps regarding technological initiatives aimed at addressing management of digital scholarly workflow in humanities scholarship.

The study was comprised of two research phases, each of which focused on a specific set of research questions and goals. The first phase included a web-based survey that consisted of twenty-five questions, which, in addition to demographic information, included queries about data searching, storing, citing, sharing, and archiving practices, as well as about scholars' experiences in using digital research tools and resources. A total of 196 faculty (59% female / 41% male) completed the survey, most of them tenured faculty, with fixed-term (non-tenure track) faculty, and tenure-track faculty following. The Humanities tended to have older respondents (over 40 years of age), while the sciences and social sciences faculty skewed lower in age.

The second phase of the study included a set of face-to-face ethnographic interviews. A total of twenty-three scholars volunteered to participate in the interviews, and they were

equally divided along the lines of disciplinary profiles, academic ranks, and gender: 13 were faculty in the humanities and social sciences (HSS) and 10 in the sciences; 11 were tenure-track and 12 tenured faculty; 13 were female and 10 men. The interviews were semi-structured and, on average, lasted an hour. The interviewees were audio-recorded and then transcribed by a professional transcriptionist. The interview transcripts were first coded into broader categories (nodes) by two independent coders.

We then proceeded with focused coding, where the categories into which the data were originally coded had additionally been refined for relevant patterns, themes, and topics.

The results of our study show that digital technologies have different roles and levels of integration at various phases of scholarly workflow. For instance, digital tools are actively used for finding, storing, and archiving research materials. This finding is true across disciplines, although certain disciplinary differences can be traced. For instance, while the majority of respondents across disciplines (92%) actively store research materials important to them, humanities scholars reported the highest percentage of lost and inaccessible research files; predominantly (27%), inaccessible files resulted from failing to migrate research materials from obsolete to contemporary digital formats. Similarly, while searching for information electronically is a standard, daily practice of our respondents regardless of their disciplinary background and/or level of technical proficiency, humanities scholars commonly prioritize the Penn State library catalog as their search and access points, while scholars in the sciences prioritize Google Scholar. Our results also show, however, that across disciplines, the path towards finding information commonly starts with Google Search and Google Scholar, especially for scholars engaged in discovery search, which reaffirms results of other recent studies indicating the increasing prevalence of commercial over academic services for scholars' information search (see: Nicholas et al., 2011; Kortekaas, 2012)

The results of our study further show that, in the phases of data collecting and analysis, the use of digital technologies significantly differs across disciplines. Our respondents in the science commonly noted that their work would be impossible without digital technologies, and scholars in the social sciences indicated digital tools and methods becoming 'a new normal' in their data gathering and analysis practice. Contrary to this, respondents in the humanities, with a few exceptions, implied the lack of digital technology use in those phases of their research process. Parallel with this, however, they indicated awareness of digital tools and methods that could facilitate their analytical practice, suggesting the lack of available training and time as key impediments to developing literacies needed for mastering those tools.

Disciplinary differences were evident in the activities of data sharing and communication, particularly in the use of social media. With regard to data sharing, two thirds (63%) of scholars in the sciences indicated that they actively share their research data, while a nearly identical percentage of the humanities scholars (69%) indicated opposite practice. Yet we found that in addition to disciplinary differences, differences in academic standing also influence data sharing practices of our respondents, with tenure-track faculty being more protective of their data than tenured scholars. We further observed widespread use of digital technologies in scholarly communication across disciplines, with a noticeable difference being frequent social media use among the humanities scholars, and nearly non-existent use among respondents in the sciences.

Annotating and reflecting emerged as research phases where the use of digital technologies is most idiosyncratic, that is, based on scholars' personal preferences rather than the level of technical skills or availability of digital tools. With regard to citation, the use of citation management programs was somewhat higher in the sciences than in the HSS (55 % vs. 30 %), but the overall level of digital technology use in this research activity was lower than in other phases of the research workflow.

Conceptually, our results illustrate various ways in which integration of digital tools in one phase of the research

processes influences other segments of the workflow. For example, scholars' full reorientation on electronic search and access produces an abundance of collected materials, requiring adjustments in researchers' storing, organizing, and archiving practices. As some of our respondents observed, integration of digital tools into their search activities resulted in a complete breakdown of their systems for organizing information, developed for print-based materials. Therefore, while implementation of digital tools into one phase of the workflow might be rewarding, it might also become a challenge in other phases of the workflow. This is particularly relevant in the perspective of tool development, implying that digital research tools should be designed to support a continuous research workflow instead of separate and disconnected activities.

Our findings also suggest that in a workflow of a digital scholar technical rather than traditional methodological expertise shapes interconnectedness among phases of the workflow. In our study, greater level of workflow interconnectedness was observed among scholars in the sciences, who tend to be more technologically savvy than scholars in the humanities and social sciences. This, as well as our previously mentioned study findings, indicates a significant scope of disciplinary differences with regard to the use of digital technologies in scholarly work. Broadly conceived, these disciplinary differences can be conceptualized as inherent and acquired. As an example of inherent disciplinary differences we could understand data privacy requirements, which widely differ across disciplines and, as our findings show, significantly determine the type and level of digital technology use. Acquired differences on the other hand can be observed in a set of habits and assumptions rooted in a particular community of practice. Technical architecture of digital research tools needs to support specific disciplinary needs in ways that address both inherent and acquired disciplinary differences. Data storage and management, for instance, has been identified as a dire problem across disciplines, but with distinctive disciplinary needs.

The next phase of our study (2014-2016) will be devoted to developing a digital research tool for humanities scholarship using *Zotero* as a test platform, in collaboration with George Mason University. Based on the results of the first phase of our study, we will focus on unifying several phases of the research workflow, and facilitating elements such as better integration of finding and archiving into the scholar's online path. Discovery must be better finessed for the end user, and search and retrieval should be fully integrated into an interface that also allows annotation, organization, and archiving of research materials. Also, since the loss of information among the humanities scholars is significant, there is a need to build into the research workflow easy strategies for users to self-archive their work in storage services that are inherent to the individual or the institution. Optimizations to connect the institutional repository within *Zotero*, as well as expose references and metadata within uploaded PDFs will be explored.

Building a multi-dimensional space for the analysis of European Integration Treaties. An XML-TEI scenario

Armaselu, Florentina

florentina.armaselu@cvce.eu
CVCE

Allemand, Frédéric

frédéric.allemand@cvce.eu
CVCE

Goal of the project

The European Union (EU) is a “union of law”: it is created by law; it enacts laws which confer rights upon EU citizens and impose duties; it acts in accordance with its law and under the legal review of the Court of Justice. The EU has, by its own, neither peoples nor territories. Its existence is entirely enclosed in its ability to bring closer its members states as well as the Europeans through the law. Thus, the knowledge (and the understanding) of the EU law is an essential part of an ongoing democratic process. In practice, pure and perfect knowledge of the law has always faced many difficulties. This is particularly true for the EU constitutional law. It is an economic and technical law faced with significant several changes in recent years and published in all official languages of the Union (24).

The goal of the project is to create a tool for assisting the researchers in European Integration Studies in the analysis of the EU treaties. The system should allow the user to navigate through the treaties along with different axes of inquiry: access to a particular unit inside a treaty (part, section, chapter, article, etc.); multilingual alignment; modifications operated upon or by a treaty and the history of its different (consolidated) versions; status (entered into forced, repealed); possibility to add and retrieve user's comments. The tool aims also to provide the EU citizens with the consolidated and the original versions of EU treaties enriched with additional materials (i.e. contextual resources, legal & economic doctrine, case law, ...). The documents under study are part of the CVCE's Lisbon research corpus , including founding treaties of the European Union and the treaties modifying them.

Although Web-based services allowing the navigation (EUR-Lex¹, LegiFrance², DOCLEG³, the versioning (Progilex⁴, MetaLex⁵) and/or annotation (AT4AM for All⁶) of legal documents already exists, there is no integrated solution addressing all the questions entailed by our research. So far, our experiments have been dealing with the identification of the documents structure and their relations, and the construction of small prototypes using XML-TEI as an encoding format. As the project is still in an exploratory phase, the paper will focus on the theoretical bases of the project, first experimental results and further development.

Overview of the process of creating/modifying European Integration Treaties

The EU – formerly the European Communities – was established by the Rome Treaty concluded in March 1957. Since then, this treaty was modified more than 20 times, either in application of the general revision procedure or by simplified revision procedures. Every revision is enacted in the form of a legal act – be it a new treaty such as the Maastricht treaty (1992) or a secondary law like a Council decision. The act which introduces changes exists by itself, in addition to the act which is modified. However, nowhere the consolidated versions of the treaties and all their modifications are provided, even by Eur-Lex – the web service maintained by the Publications Office of the EU. This makes any analysis of EU legal texts highly tricky as any user looking for an updated version of a specific text has to compare its original version with all its subsequent revisions. Due to their complexity and their multi-linguistic nature, EU legal texts require also corrections which are legally binding. The rule relating to the allocation of seats in the European Parliament among the EU Member States since 1951 is illustrative of the complexity of the EU law. The modifications are laid down either in primary law or in secondary law; they insert, repeal, include either a whole Article or a part of it. The dates of adoption, of publication, of entry into force, of effective implementation vary from one modification to another. Some modifications were published but never entered into force applied, some other were changed before they become effective.

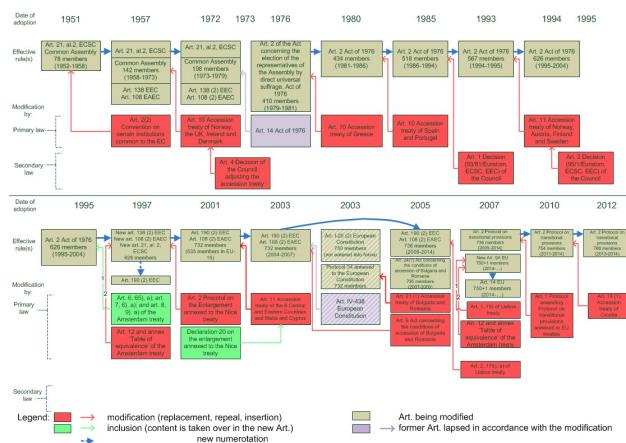


Fig. 1: Complexity of the EU treaties revision process: the example of the provisions on the allocation of seats in the European Parliament from 1951 to 2013.

Documents structure. Relationships

Structure of the documents

The text of a treaty can include the following elements:

- preamble;
- main body of the treaty;
- annexes;
- protocols;
- declarations;
- final acts;
- corrections.

The corresponding TEI document may contain, apart from the metadata encoded in the *TEI Header*⁷, a *text element*⁸ with the following constituents: *front*(title, preamble); *body* (main body of the treaty); *back* (protocols, declarations, final act, corrections). Except from the title and the main body of the treaty, the other elements are optional. Imbrication between some of these components is possible. For instance, protocols or declarations can appear independently or be included in a final act. The main body of a treaty is structured in Fig. 2. Other components, like protocols or annexes, may have a similar configuration.

Main body of the treaty/protocol/annex ...

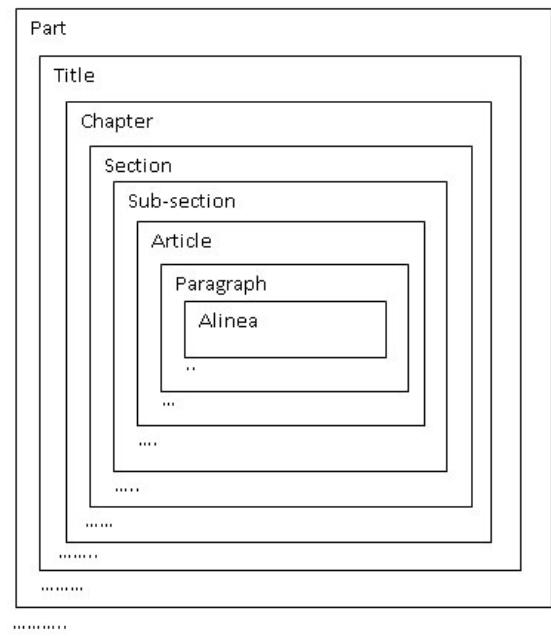


Fig. 2: Structure of the main body of a treaty

Smaller items can be further identified but the lowest unit considered in the study is the alinea (it corresponds to a paragraph in the non-legal writing)⁹.

Relations

The relations to be modeled operate either between treaties (Fig. 3) or at the fragment level, for example, an article from a treaty is modifying another article from a different treaty (Fig. 4).

4). The relations between treaties (see also¹⁰) are represented below and correspondingly in Fig.3.

- amended_by / amends (oblique arrows pair);
- previous /next versions (other than linguistic) (horizontal arrows);
- linguistic_version (oblique arrow).

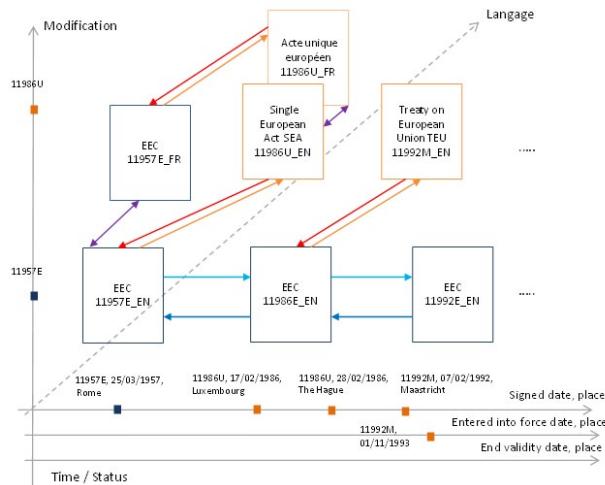


Fig. 3: Inter-treaties relations

The figure shows, along with the Modification and Time/Status axes, how the Treaty establishing European Economic Community (EEC, 1957) is amended by the Single European Act (SEA, 1986) and by the Treaty on the European Union (TEU, 1992), two subsequent consolidated versions being produced accordingly (EEC, 1986; EEC, 1992) (numeric codes are inspired by¹¹). The Language axis adds another dimension to the representation of the different linguistic versions of the treaties. Since the relations actually produce a multi-dimensional space, we can imagine the representation as functioning by parallel planes, rather than in a single three dimensional reference system. Moreover, the Time/Status axes are used to define three different timelines, for the creation/signature, entered into force and end of validity dates and places.

The relations at the fragment level can be of the following types:

- modified_by / modifies;
- cited_by / cites;
- previous / next.

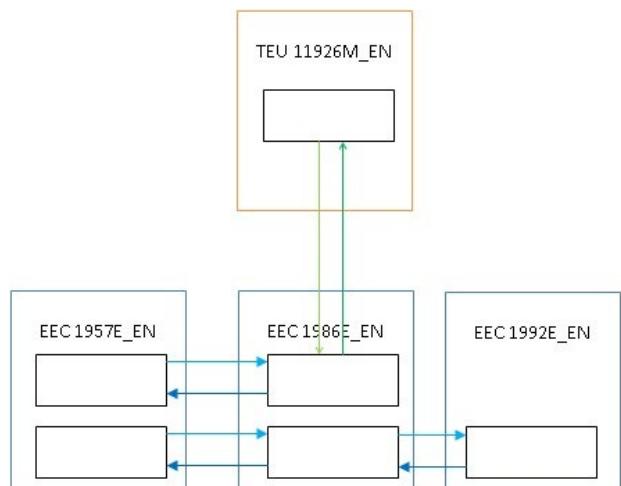


Fig. 4: Relations at the fragment level

Fig. 4 illustrates the previous/next relations (horizontal arrows) between fragments (dispositions) from a version to the other and how a disposition from TEU (1992) repeals another from EEC (1986) (vertical arrows).

Experiments

The experiments conducted so far dealt with encoding the structure of the treaties main body (Fig. 2) and the multilingual alignment. The production of XML-TEI documents involved a transformation chain represented in Fig. 5.

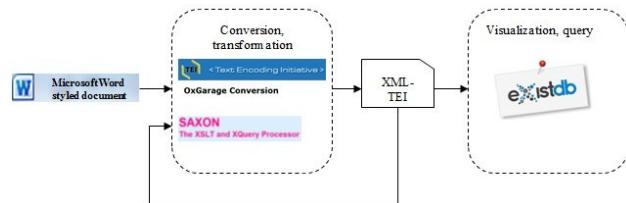


Fig. 5: XML-TEI transformation chain

First, the styled Microsoft Word documents, with styles corresponding to the structural components down to the level of Article (Fig. 2), were converted into XML-TEI (P5) using the OxGarage converter¹². The components were encoded using div elements. An XSL¹³ file was created in order to enrich the encoding produced by the first conversion with attributes (@typeaccepting as values the components names, @xml:id, @n) for every div element. The transformation performed via Saxon¹⁴ also included procedures for the delimitation and identification of paragraphs and alineas (not marked by Microsoft Word styles). The resulted XML-TEI files were stored in an eXist-db¹⁵ database and HTML, CSS and XQuery¹⁶ scenarios were added for visualization and queries. Fig. 6 shows the result of a search at the level of alinea and its multilingual alignment.

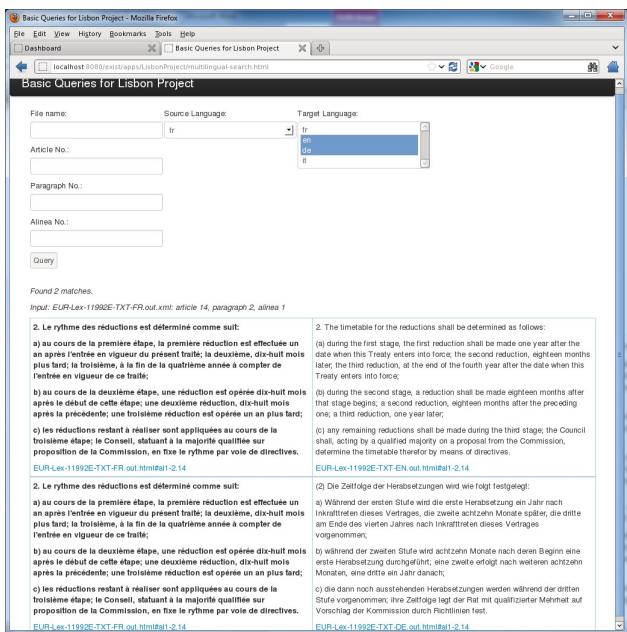


Fig. 6: Search by alinea. Multilingual alignment

Further work

The complexity of the relations modeling involved in our study mainly resides, on one hand, in their temporal dimension (the modifications, the generation of new versions, their validity, and implicitly their relations, query and retrieval should operate in terms of time points and time frames) and, on the other hand, in the linguistic diversity of their expression (multilingual nature but also different ways of expressing, inside the same language, how a given fragment from a treaty modifies another). Our current inquiry is therefore related to the expression of the: types of amendment (e.g. *repeal*, *insertion*, *substitution*)¹⁷; nature of the reference (*document/fragment*, *internal/external*, *simple/multiple*, *contextual/non-contextual*)¹⁸; *active/inactive* time intervals¹⁹. Other elements under study and experimentation are the relations encoding (TEI linking specifications²⁰; xLink, xPointer^{21, 22}), as well as the potential use of TEI extensions for legal texts²³. Aspects related to the balance between manual versus automatic processing, the corresponding workflow, and the maintenance strategies allowing the incorporation of new data or the integration of user's intervention are also to be considered.

The presentation will focus on the theoretical bases of the project and the experimental results.

References

1. **EUR-Lex Access to European Union**, eur-lex.europa.eu/en/index.htm.
2. **Legifrance, le service public de la diffusion du droit**, www.legifrance.gouv.fr/.
3. **Documents DOCLEG**, www.riziv.fgov.be/webprd/docleg/cgi-bin/cgint.exe?9&ulang=fr.
4. **Progilex**, legal publisher in Luxembourg, www.legitech.lu/fr/progilex/presentation.
5. **CEN MetaLex**, Open XML Interchange Format for Legal and Legislative Resources, www.metalex.eu/.
6. **AT4AM for all**, the web-based amendment authoring tool used at the European Parliament, www.at4am.org/.
7. **Text Encoding Initiative, P5, TEI Header**, www.tei-c.org/release/doc/tei-p5-doc/en/html/HD.html.
8. **Text Encoding Initiative, P5, Default Text Structure**, www.tei-c.org/release/doc/tei-p5-doc/en/html/DS.html.
9. **Mencia, Eneldo Loza** (2009), *Segmentation of legal documents*, Proceedings of the 12th International Conference on Artificial Intelligence and Law, Barcelona, Spain, p. 88–97.

10. **Treaty on European Union**, 11992M/TXT, Eur_Lex, eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:11992M/TXT:EN:NOT#top.

11. **HOW CELEX NUMBERS ARE COMPOSED**,

EUR-Lex, 2011, 2013, eur-lex.europa.eu/en/tools/HowCelexNumbersAreComposed.pdf.

12. **OxGarage Conversion**, www.tei-c.org/oxgarage.

13. **The Extensible Stylesheet Language Family (XSL)**, www.w3.org/Style/XSL.

14. **Saxon**, The XSLT and XQuery Processor, saxon.sourceforge.net.

15. **eXist-db Open Source Native XML Database**, exist-db.org.

16. **W3C XML Query (XQuery)**, www.w3.org/XML/Query/.

17. **Spinosa, P., Giardiello, G., Cherubini, M.** (2009), *NLP-based metadata extraction for legal text consolidation*, ICAIL '09 Proceedings of the 12th International Conference on Artificial Intelligence and Law, pp. 40-49

18. **Martinez Gonzalez et al.** (1997), *Electronic Manipulation of European Texts about Conflicts of Jurisdiction : a Semantic Web Tool*, Paper delivered at the Text Encoding Initiative Tenth Anniversary User Conference, November 1997, www.lefis.org/images/documents/outcomes/lefis_series/lefis_series_2/capitulo4.pdf.

19. **Boer, A., Hoekstra, R., Winkels, R.** (2002), *METALEx: Legislation in XML*, in T. Bench-Capon, A. Dascalopulu and R. Winkels (eds.), *Legal Knowledge and Information Systems, Jurix 2002: The Fifteenth Annual Conference*, Amsterdam: IOS Press, pp. 1-10, citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.12.3189&rep=rep1&type=pdf.

20. **Text Encoding Initiative**, P5, Linking, Segmentation, and Alignment, www.tei-c.org/release/doc/tei-p5-doc/en/html/SA.html.

21. **W3C XML Pointer**, XML Base and XML Linking, www.w3.org/XML/Linking.

22. **Martinez M.M., De La Fuente, P., Derniame, J.-C., Pedrero, A.** (2003), *Relationship-based dynamic versioning of evolving legal documents*, INAP'01, Proceedings of the Applications of prolog 14th international conference on Web knowledge management and decision support, Springer-Verlag Berlin, Heidelberg, pp. 290-305.

23. **Finke, N.D.** (1997), *TEI Extensions for Legal Text*, *Text Encoding Initiative Tenth Anniversary User Conference*, www.stg.brown.edu/conferences/tei10/tei10.papers/finke.html#fn0.

The Layered Text. From Textual Zoom, Text Network Analysis and Text Summarisation to a Layered Interpretation of Meaning

Armaselu, Florentina

florentinaa@zoomimagine.com
Zoomimagine

Introduction

In her article describing a night performance with a dancers troupe¹, Rambo Ronai proposes a "layered account" combining different perspectives, that of a dancer, wrestler, ethnographer and writer reflecting upon Derrida's concepts of "mimesis" and "under erasure", and the metaphor of drawing/writing as a way to express the "layering process" of live experience. Starting also from ethnographical observation and Ryle's notion of "thick description", Geertz² considers the interpretation process as founded upon "piled-up structures of inference and implication" and the detection in the observed object of a "stratified hierarchy of meaningful structures", like the "twitches, winks, fake-winks, parodies, rehearsals", etc. in the twitching/

winking scenario or the Jewish, Berber and French "frames of interpretation" in Cohen's story.

Our proposal relies on the hypothesis that a "layered" representation of an electronic text can bring into light some aspects related to the production and circulation of meaning in the reading/interpretation and writing process. The study refers to models and methods like textual zoom, text network analysis and text summarisation and proposes a combined approach for structuring the text on "layers of meaning".

Textual zoom, z-text

The model of *z-text*³ was inspired by Neal Stephenson's⁴ fictional construct, a primer whose content expands itself in its interaction with the reader. A *z-textual* layout supposes a hierarchical structure of *z-lexias* (after Barthes' *lexia*⁵, a unit of reading and analysis), i.e. potentially extensible units, disposed on levels of detail, along with the Z-axis. The processes of reading and writing *z-lexias* are called *z-reading* and *z-writing*. A parent-*z-lexia* consists in a fragment which has engendered descendants, i.e. has been expanded on the subsequent levels of the representation, like *z/1* and *z/3* in Fig. 1. Each *zoom-in* operation performed by the reader replaces a *z-lexia* visible on the roll (the display device on the topmost plane) with its next-level children (if any), while a *zoom-out* substitutes all the displayed children with their previous-level parent (if any). The zooming mechanism entails a back and forth movement through the layers of text and the dynamic projection of *z-lexias* on the displaying device.

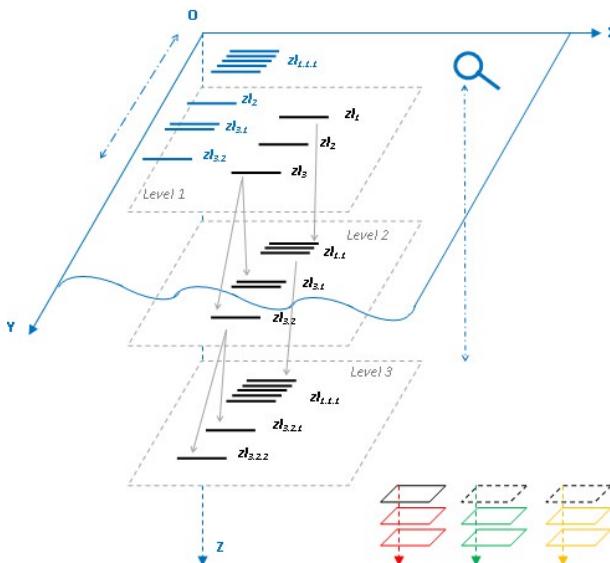


Fig. 1: Z-text model

The "tri-dimensional space" of a *z-text* can be turned multidimensional if the same *z-lexia* is expanded and then explored by different types of magnifying glass, i.e. following different points of view or perspectives (sometimes contradictory, like in Pavić's Khazar controversy viewed through the lens of the red, green or yellow books).

Figure 2 presents a z-textual layout of Barthes's *S/Z*. The representation starts with a fragment from Balzac's *Sarrasine* on the first level. New details are added gradually on each level:

- the "units of reading"(lexias) and the attached interpretation codes, HER, SYM, SEM, etc. (level 2);
 - the description of the interpretation method and its codes, as a way to understand the plurality of text defined as a "galaxy of signifiers" (level3);
 - more insight into the "step by step" analysis of text "working back along the threads of meanings", the "weaving voices" made apparent by the five codes, and the evaluation process echoing the writing practice and allowing us to distinguish the "readerly" and the "writerly" (level 4);
 - emphasis on the idea of enclosing the text in a fixed structure versus providing it with a "structuration", on

considering the text as a process rather than a product, and on the reversibility of the writerly text, proved in the example by actually turning Balzac's Sarrasine into Barthes's *S/Z* (level 5).

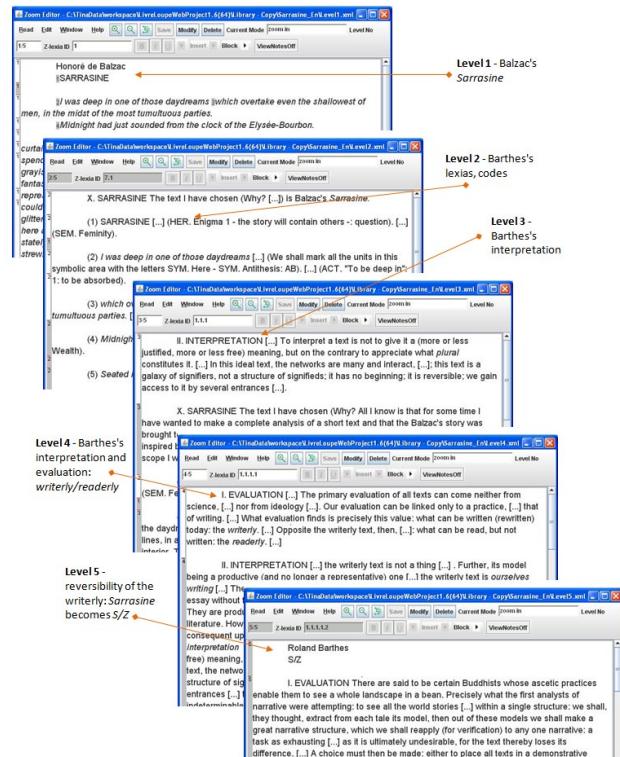


Fig. 2: Layers of meanings in Barthes's S/Z

The process can continue with the reader's interpretation on Barthes's interpretation of Balzac, the *zoom-in*, *zoom-out* mechanism allowing to move back and forth from the initial text to an interpretation (or interpretation of interpretation, ...) of it, through different layers of meaning involving variable degrees of details.

Text network analysis

The z-textual layout of S/Z was based on the assumption that Barthes's analysis contains in itself a certain stratification on levels of signification that can allow the gradual transformation of one text into the other. The levels texts were made up by fragments, not necessarily contiguous as in their original form, but following a certain hierarchical logic (e.g., level 1 - *Sarrasine*; level 2 - SEM, HER, ...; level 3 - codes explained, etc.).

Further analysis consisted in the use of *TexTexture*⁷, a visualisation tool based on the concepts of text network analysis and betweenness centrality⁸. The five files, corresponding to z-text levels (Fig.3), were processed via the *TexTexture* online service, each file representing a gradual enrichment of the *Sarrasine* text with Barthes's analysis as described above.

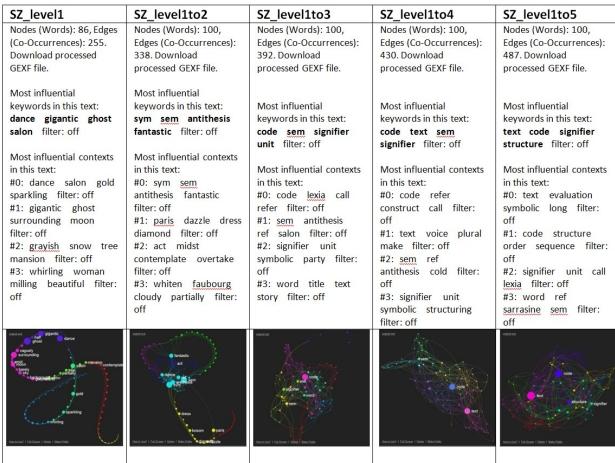


Fig. 3: TexTexture. S/Z z-textual layout

Figure 3 shows the most influential concepts in the texts, i.e. the words with the highest value of betweenness centrality (measuring how often a node appears between any other nodes in the network). The most influential contexts are also displayed as determined by nodes with high degree of connectivity (number of edges). While high degree nodes can be influential within a given context, high values of betweenness centrality characterize words supposed to function as shifting points between different clusters of meaning.

A left to right scan of the five columns denotes a certain dynamics in the transition from the Balzacian text to its Barthesian interpretation. Thus, from a first level description of the party *salon*, the emphasis shifts to a more analytic perspective articulated around the interpretation codes (*sem*, *act*, *sym*) and terms like *antithesis* and *fantastic*, resonant with the already highlighted *ghost*, *dance*, *moon* on the previous level. More details on the interpretation procedure added on the third, fourth and fifth levels bring about meaning circulation through nodes like *code*, *signifier*, *unit*, *lexia* (level1 to 3) through *text*, *voice*, *structuring*, *signifier* (level1 to 4) and finally to *code*, *text*, *signifier*, *structure*, *evaluation* (level 1 to 5), elements that seem to approximate the different layers of meaning embedded in the z-textual layout.

Text summarisation

Text summarisation, extractive⁹ or abstractive¹⁰ techniques, represents another point of interest for our inquiry. For instance, studies like^{11, 12, 13, 14} making use of graph-based models in order to encode the structure of a text and to compute the most salient sentences/fragments to be included in the summary, or tools for variable summarisation, like^{15, 16} allowing to generate summaries covering a given percentage from the initial text.

Towards a layered interpretation of meaning

Our proposal consists in a theoretical approach combining textual zoom, text (network) analysis and gradual summarisation for the detection of key elements and the representation of layers of meanings in a text. The combined construct, at this point defined only at a conceptual level, may enclose potential functionalities such as:

- highlight the most influential concepts/contexts in the text and their ranks computed according to particular relevance criteria (e.g. betweenness centrality or other);
- propose candidate sentences to be included on different levels of summarisation, possibly based on a certain rank order (for instance, starting with lower or higher rank constituents on the lower levels);
- assist the user in building further summarisation levels by gradually adding remaining constituents, until the whole text is covered;

- integrate the levels of summarization into a z-text layout that can be eventually explored by *zoom-in* and *zoom-out*.

The structure can be considered (in a kind of "deformative" interpretation¹⁷ or close/distant reading scenario) in order to explore the layers of meanings "hidden" in a text. Our hypothesis is that words and sentences may appear in disparate places throughout the text, but from an interpretative or writing perspective they may belong to the same conceptual or symbolic level. Grouping these fragments on layers of meaning may bring new light on the process of text production and understanding.

Similar with the S/Z experiment, we may imagine, for instance, a "step by step" passage from the analyzed text to deeper analytic levels (like in Auerbach's¹⁸ reflections on the representation of reality in Western literature, starting from a close reading of *Odysseus's Scar*, or in Greenblatt's¹⁹ new-historicist analysis of *Midsummer* leading to a reconstruction of the historical-cultural context having inspired it). Other examples can deal with the variable degree of proximity/ distance of the reader to a textual object (as in Shakespeare's *Venus and Adonis* or in a hypothetical "behaviorist" narrative progressively enriched with characters' psychology), the gradual movement from simple to complex, intuitive to abstract in pedagogical or philosophical scenarios (e.g. Wittgenstein's²⁰ *Tractatus*), as well as the "layered" representation of a growing text - from a few initial paragraphs to a full-fledged story, resulted from a writing process.

A layered interpretation of meaning may be aligned, besides Rambo Ronai's and Geertz's theses, with Iser's²¹ assumption on the process of "anticipation and retrospection" implied by the act of reading , and Schor's²² absorbed (or absent) detail and its capacity to "persist and inform in absence". Every layer of meaning carries a potential both for retrospection and anticipation, acting, in an absence/presence scenario, as a bridge between the already-said and what is about to be articulated.

The presentation will include both theoretical aspects and a demo on the proposed topic.

References

1. **Rambo Ronai, C.** (1999), *The Next Night Sous Rature: Wrestling With Derrida's Mimesis*, Qualitative Inquiry 1999 5: 114, pp. 114-128.
2. **Geertz, C.** (1973), *The Interpretation of Cultures*, Basic Books, New York, pp. 6-9.
3. **Vasilescu (Armaselu), F.** (2010), *Le livre sous la loupe : Nouvelles formes d'écriture électronique*, Ph.D. Thesis, Papyrus, University of Montreal Institutional Repository, papyrus.bib.umontreal.ca/xmlui/handle/1866/3964 ;jsessionid=5DEBDCDB0FDA32C06644880B79A9B941.
4. **Stephenson, N.** (2003), *The Diamond Age or A Young Lady's Illustrated Primer*, Bantam Books, New York, 1995, new editions 1996, 2000, 2003.
5. **Barthes, R.** (1974), *S/Z*, first edition 1970, translated by Richard Miller, Hill and Wang, New York.
6. **Pavić, M.** (1988), *Dictionary of the Khazars*. A lexicon novel. Female Edition, Vintage International, New York.
7. TexTexture, visualize any text as a network, textexture.com .
8. **Paranyushkin, D.** (2011), *Identifying the Pathways for Meaning Circulation using Text Network Analysis*, Published in Nodus Labs, December.
9. **Gupta, V. and Lehal, G. S.** (2010), *A Survey of Text Summarization Extractive Techniques*, Journal of emerging technologies in web intelligence, vol. 2, no. 3, august 2010.
10. **Genest, P-E., Lapalme, G.** (2011), *Framework for Abstractive Summarization using Text-to-Text Generation*, Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, Portland, Oregon, 24 June 2011, pp. 64-73, aclweb.org/anthology//W/W11/W11-1608.pdf .
11. **Mihalcea, R.** (2005), *Language Independent Extractive Summarization*, Proceedings of the ACL Interactive Poster and Demonstration Sessions, pages 49–52, Ann Arbor, June 2005. c2005 Association for Computational Linguistics.

12. Mihalcea, R. and Tarau, P. (2004), *TextRank – bringing order into texts*. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2004), Barcelona, Spain.
13. Barzilay, R. and Elhadad, M. (1997), *Using Lexical Chains for Text Summarization*, In Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization.
14. Ganesan , K., Zhai, C.X., Han, J. (2010), *Opinosis: A Graph-Based Approach to Abstractive Summarization of Highly Redundant Opinions*, Proceedings of the 23rd International Conference on Computational Linguistics (Coling '10), sifaka.cs.uiuc.edu/czhai/pub/coling10-opinosis.pdf .
15. Text Compactor, Free Online Automatic Text Summarization Tool, textcompactor.com .
16. Microsoft Word, AutoSummary Tools.
17. McGann, J. (2004), *Radiant Textuality: Literature after the World Wide Web*, Palgrave Macmillan.
18. Auerbach, E. (2003), *Mimesis. The Representation of Reality in Western Literature*. Princeton University Press, Princeton and Oxford, 1953, 2003.
19. Greenblatt, S. (2004), *Will in the World*. How Shakespeare became Shakespeare, New York–London, W.W. Norton & Company.
20. Wittgenstein, L. (1999), *Tractatus Logico-Philosophicus*, Translated by C.K. Ogden, Dover Publications, Inc., Mineola, New York.
21. Iser, W. (1974), *The Implied Reader*, The John Hopkins University Press, Baltimore and London.
22. Schor, N. (2007), *Reading in Detail: Aesthetics and the Feminine*, first edition, 1987, Routledge, New York, London, p. 72.

Binarization-free Text Line Extraction for Historical Manuscripts

Arvanitopoulos Darginis, Nikolaos

nick.arvanitopoulos@epfl.ch
EPFL, Switzerland

Süsstrunk, Sabine
EPFL, Switzerland

1 Introduction

Nowadays, large collections of old historical manuscripts, which contain valuable information about our cultural heritage, exist in libraries around the world. Recently, there has been much interest in their digitization for preservation reasons, since many of the available manuscripts' quality has deteriorated from exposure to the environment. Digitization though is only the first step to make the information contained in manuscripts accessible to researchers and to the interested public. What we create after digitization is only a "digital image" of the page and further processing steps need to be applied during the *handwriting recognition process*, so that the manuscript's content is transformed into a form that is interpretable by a computer.

One important step in the handwriting recognition process is that of *text line extraction*, which aims at extracting individual text lines from the manuscript page. In this paper, we propose a binarization-free text line extraction method using seam carving. The main idea is to compute an energy map of the input text blocks and determine minimum energy paths that pass through them. The energy map is constructed in a way so that gaps between text lines have low energy values. Therefore, a minimum energy path will pass only through these regions and will successfully separate two text lines.

Our algorithm has the following two advantages:

1. We make direct use of the original image representation of the manuscript page without any need for prior binarization, which can introduce information loss. This loss can produce

unreliable results for the text line extraction algorithm (see Figure 1).

2. Our algorithm is general and can be applied to diverse manuscripts of different time periods and handwritings. Results in Figures 3 and 4 show the applicability of our algorithm to diverse historical manuscripts.

2 Related Work

We briefly summarize the research that has already been done for text line extraction.

Most of the state-of-the-art approaches operate on a binary image of the historical manuscript. One method based on dynamic programming computes the paths with minimum cost between two consecutive lines [11] and has been extensively used in automatic transcription and ground truth creation of historical documents [8, 7]. The work of [6] is based on horizontal projection profiles of black pixel changes. An additional post-processing step is applied, which follows the contour of the ink obtaining curve-linear line separators. Another similar approach is proposed in [13] where the output of the horizontal projections is post-processed based on properties of the computed connected components. The works of [12, 10] are based on the Hough transform, which is able to detect straight lines in images. Smearing methods, such as the ones in [17, 16, 14], try to fill-in the white pixel gaps with black pixels if their distance is less than a threshold. That way, homogeneous blocks of the document page are grouped together. Other approaches use multi-oriented filters and active contours for text line extraction [4, 5].



(a) Gautier de Metz, *Le miroir du monde*. Copy from François Buffereau, secretary of Antoine de Gingins (1475-1500).
(b) Gautier de Metz, *Le miroir du monde*. Copy from François Buffereau, secretary of Antoine de Gingins (1475-1500). Binarized version.

Fig. 1: Left: seams generated using the original scanned image of the manuscript as input. Right: seams generated using a binary version of the original manuscript scan as input. The information loss in the binary version is so extensive that the generated seams do not clearly separate text lines of the original manuscript scan.

A notable exception of an algorithm, which does not depend on binarization, is the work of , where the text lines are found using extracted features from interest points of the original manuscript image. A very recent work uses a framework similar to that in adapted this time to the text line extraction problem.

Our method is closely related to the one in, where the authors use seam carving to generate seams that pass through connected components of a binary image. Unclassified components, which do not belong to any text line are assigned in a post-processing step according to their position and geometric characteristics. The main difference in our approach is that we do not need to binarize the input, which can lead to information loss. Additionally, we are able to generate robust text contours even for manuscripts of deteriorated quality (see Figures 1, 4). The text contours can always be overlaid on the original manuscript scan, even if they have been generated using as input a binarized version of the original scan. However, the technique of , which assigns text components to lines is not able to extract lines from the original manuscript, since the binarization process is not reversible. Binary text components

contain only a subset of the information available in the original manuscript image.

3 Our Approach

Our proposed algorithm is inspired by *Seam Carving*, a computer vision algorithm used for image resizing [1]. We build upon this idea and propose a seam carving algorithm, which operates on the original color image and extracts lines in a sequential way. First, an energy map is calculated and the minimum energy path is computed based on dynamic programming. From the peaks of the horizontal projection profile of the derivative image we can find horizontal line positions. In each such region between two consecutive lines, we apply our seam carving algorithm sequentially until the whole manuscript image has been processed. In Figure 2 we show some examples of seams between two such lines.

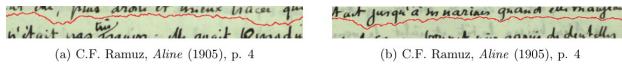


Fig. 2: Examples of image blocks and their computed seams.

In the following we use the convention that an image $I \in \mathbb{R}^{n \times m}$ converted to grayscale has n rows and m columns. The notation $I(i,j)$ denotes the image value at the i -th row and j -th column. The coordinate system has its origin in the upper left corner of the image.

3.1 Energy Map

We modify the energy function proposed in [1] so that it can be effectively used for generating text line separation contours. First, we compute an edge image as

$$E(i, j) = \left| \frac{I(i, j+1) - I(i, j-1)}{2} \right| + \left| \frac{I(i+1, j) - I(i-1, j)}{2} \right|.$$

Let us denote the energy map between two text lines by $E_b = E(J, :) \# R^{k \times m}$, where $J = \{s, \dots, e\}$ is the set of i coordinates between the start and the end of the energy block map and $l = e - s + 1$. This energy block is weighted by the following linear function, which penalizes the larger i coordinates more:

$$w(i) = \frac{1}{n-1}i + \frac{n-2}{n-1}, \quad i = 1, \dots, l, \quad w(i) \in [1, 2],$$

and the final energy map for this block is

$$E_{b,f}(i, j) = E_b(i, j) * w(i), \quad i = 1, \dots, l, \quad j = 1, \dots, m.$$

The idea behind this weighting is the observation that we want our seam to be closer to the upper line than the lower one. This will correct for situations where the author has written words in the gap between lines, which always belong to the lower line (see Figure 2a).

3.2 Seam Computation

A seam that passes horizontally through an image block can be defined as

$$\mathbf{s}_b^h = \{\mathbf{s}_{b,j}^h\}_{j=1}^m = \{(y(j), j)\}_{j=1}^m, \quad \forall j, |y(j) - y(j-1)| \leq 1, \quad y(j) = 1, \dots, l.$$

The seam computation is done using dynamic programming. We look for the optimal seam in the image block that minimizes

$$\mathbf{s}_b^* = \arg \min_{\mathbf{s}_b} \sum_{j=1}^m E_{b,f}(\mathbf{s}_{b,j}^h).$$

The first step is to traverse the image block and compute the cumulative minimum energy M_b for all possible connected seams for each pixel position:

$$M_b(i, j) = E_{b,f}(i, j) + \min(M_b(i-1, j-1), M_b(i, j-1), M_b(i+1, j-1)).$$

The minimum value of the last column in M_b will indicate the end of the minimal connected horizontal seam. Therefore, in the second step we traverse the cumulative energy M_b backwards to find the path of the optimal seam.

4 Experimental Results

We apply our algorithm to original manuscript pages of the work *Aline* by the important Swiss-French writer Charles-Ferdinand Ramuz. Some examples of manuscript pages overlaid with the text line extraction seams are shown in Figure 3. We observe that our algorithm creates seams that pass through parchment regions, successfully segmenting the text lines. Even when the writer corrects a line or a word and writes above, the seam is able to avoid cutting the text and assigns the word to the line below it. In order to illustrate the ability of the algorithm to generalize to diverse manuscripts, we provide in Figure 4 results on manuscripts of the 16-th and 18-th century. We observe that our algorithm can be applied to manuscripts of very different quality and handwriting styles.

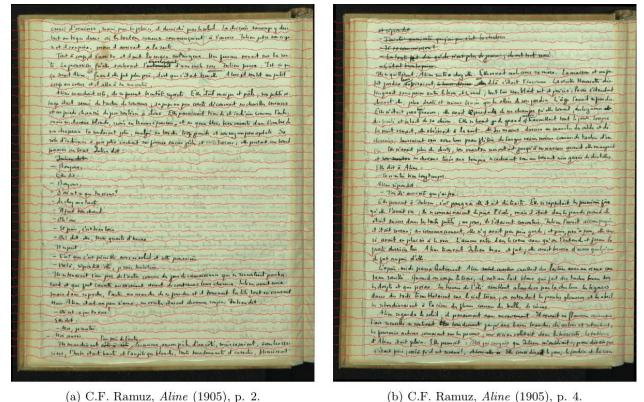
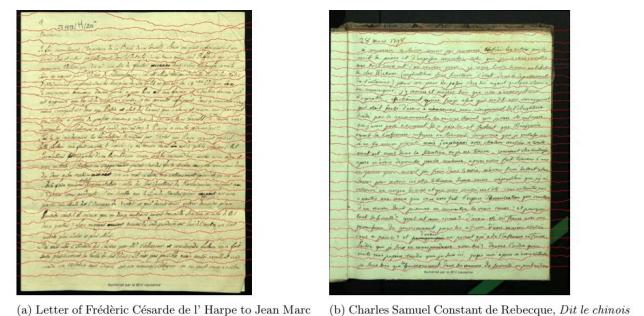
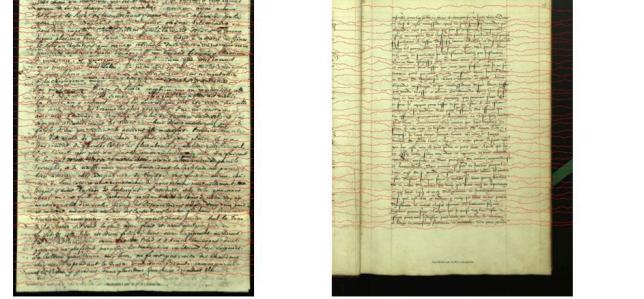


Fig. 9: Seam carving results on two pages of *Aline* (1905), C.F. Ramuz.



(a) Letter of Frédéric Césarde de l' Harpe to Jean Marc Louis Favre (1774), page 1.

(b) Charles Samuel Constant de Rebecque, *Dit le chinois* (1798).



(c) Letter of Jean Marc Louis Favre to Frédéric Césarde de l' Harpe (1782), page 1.

(d) Guillaume Budé, *Commentaire et mémorial au roi François Ier* (1522).

Fig. 10: Seam carving results on manuscripts of the 16-th and 18-th century respectively. Even in the lower left manuscript with extreme bleed-through, our algorithm is able to produce a robust result.

5 Conclusion

We propose a novel text line extraction algorithm for color scans of historical manuscripts based on seam carving. We show that we can obtain state-of-the-art results on these color images without any prior binarization. The next step after the text line extraction process is the application of a learning algorithm for handwritten word recognition in each extracted text line.

References

- [1] **Shai Avidan and Ariel Shamir** (2007). *Seam Carving for Content-Aware Image Resizing*. ACM Transactions on Graphics, 26(3):10
- [2] **M. Baechler and R. Ingold** (2011). *Multi Resolution Layout Analysis of Medieval Manuscripts Using Dynamic MLP*. In International Conference on Document Analysis and Recognition, pages 1185–1189.
- [3] **M. Baechler, M. Liwicki, and R. Ingold** (2013). *Text Line Extraction using DMLP Classifiers for Historical Manuscripts*. In International Conference on Document Analysis and Recognition.
- [4] **S.S. Bukhari, F. Shafait, and T.M. Breuel** (2009). *Script-Independent Handwritten Textlines Segmentation Using Active Contours*. In International Conference on Document Analysis and Recognition, pages 446–450, 2009.
- [5] **S.S. Bukhari, F. Shafait, and T.M. Breuel** (2011). *Text-Line Extraction Using a Convolution of Isotropic Gaussian Filter with a Set of Line Filters*. In International Conference on Document Analysis and Recognition, pages 579–583.
- [6] **M. Bulacu, R. van Koert, L. Schomaker, and T. van der Zant** (2007). *Layout Analysis of Handwritten Historical Documents for Searching the Archive of the Cabinet of the Dutch Queen*. In International Conference on Document Analysis and Recognition, pages 357–361.
- [7] **A. Fischer, E. Indermu'ile, H. Bunke, G. Viehhauser, and M. Stoltz** (2010). *Ground truth creation for hand-writing recognition in historical documents*. In IAPR International Workshop on Document Analysis Systems, pages 3–10.
- [8] **A. Fischer, M. Wuthrich, M. Liwicki, V. Frinken, H. Bunke, G. Viehhauser, and M. Stoltz** (2009). *Automatic Transcription of Handwritten Medieval Documents*. In International Conference on Virtual Systems and Multimedia, pages 137–142.
- [9] **A. Garz, A. Fischer, R. Sablatnig, and H. Bunke** (2012). *Binarization-Free Text Line Segmentation for Historical Documents Based on Interest Point Clustering*. In IAPR International Workshop on Document Analysis Systems, pages 95–99.
- [10] **L. Likforman-Sulem, A. Hanimyan, and C. Faure** (1995). *A Hough based algorithm for extracting text lines in handwritten documents*. In International Conference on Document Analysis and Recognition, volume 2, pages 774–777.
- [11] **M. Liwicki, E. Indermuhle, and H. Bunke** (2007). *On-line handwritten text line detection using dynamic programming*. In International Conference on Document Analysis and Recognition, volume 1, pages 447–451.
- [12] **G. Louloudis, B. Gatos, I. Pratikakis, and C. Halatsis** (2008). *Text line detection in handwritten documents*. Pattern Recognition, 41(12):3758–3772, dec 2008.
- [13] **U.V. Marti and H. Bunke** (2011). *Using a Statistical Language Model to Improve the Performance of an HMM-based Cursive Handwriting Recognition System*. International Journal of Pattern Recognition and Artificial Intelligence, 15(01):65–90.
- [14] **N. Nikolaou, M. Makridis, B. Gatos, N. Stamatopoulos, and N. Papamarkos** (2010). *Segmentation of historical machine-printed documents using Adaptive Run Length Smoothing and skeleton segmentation paths*. Image and Vision Computing, 28(4):590–604, April 2010.
- [15] **R. Saabni and J. El-Sana** (2011). *Language-Independent Text Lines Extraction Using Seam Carving*. In Document Analysis and Recognition (ICDAR), 2011 International Conference on, pages 563–568.
- [16] **Z. Shi and V. Govindaraju** (2004). *Line Separation for Complex Document Images Using Guuzzy Runlength*. In International Workshop on Document Image Analysis for Libraries, pages 306–312.
- [17] **K.Y. Wong and F.M** (1982). *Wahl. Document analysis system*. IBM Journal of Research and Development, 26:647–656.

Leaves of Grass: Data Animation and XML Technologies

Barney, Brett

bbarney2@unl.edu
University of Nebraska-Lincoln

Pytlik Zillig, Brian

brian.pytlikzillig@gmail.com
University of Nebraska-Lincoln

Walt Whitman's *Leaves of Grass* is one of the most famous and often-studied works of American literature. In the century since Van Wyck Brooks declared Whitman the originator of "the sense of something organic in American life"—the first to combine high art and rude experience—Whitman's masterwork has been thoroughly digested into a series of critical truisms that gives even new readers of the poems a sense of familiarity. Whether we have his poems committed to memory or have never actually read one of them, we "all know" that Whitman eschewed traditionally poetic diction, that his is a poetry of inclusiveness, that the first edition of his text in 1855 is more daring, lively, and experimental than later editions, etc.

Such axioms are comforting in the face of what is on many levels a difficult text (actually, a set of texts) to assimilate. Because Whitman applied the title "Leaves of Grass" to more than ten distinctly different volumes over the course of three and a half decades—not only adding poems but also retitling, cancelling, drastically revising, combining, and re-grouping existing ones—the goal of accurately tracing the book's evolution has consistently frustrated scholars. Recognizing that "for the reader to understand how *Leaves of Grass* grew from edition to edition, some sense had to be made of these often bewildering textual permutations," a group of late-twentieth century scholars labored for over a decade to produce a variorum edition, a tremendous accomplishment that has, unfortunately, done little to alleviate the bewilderment of permutations.

The hope that digital technologies might offer a way, at last, to lucidly represent the various stages in the evolution of *Leaves of Grass* was one of the early motivations for the creators of the *Whitman Archive* in the late 1990s. We have often revisited the question of how to convey visually the information represented in the arcane coding of the 3-volume print variorum and inherent in the separate digitized editions. Nearly two decades later, however, we haven't made much progress.

Though they cannot provide the kind of detailed, objective understanding that might be conveyed by the schematic, interactive interfaces that we've sometimes (very hazily) imagined—ones that somehow collate whole texts, poems, lines, and phrases—we have begun to experiment with distant reading strategies that provide a different sort of view. So while collation tools do not cope well with the scope of transformation involved in Whitman's reworking the first edition's 10,000-word prose preface into the 4,000-word poem "By Blue Ontario's Shore," text analysis tools such as Voyant offer a number of potentially enlightening prospects on the two works and their relationship.

Likewise, such tools can begin to offer ways to assay and quantify some of the critical commonplaces that have grown up around *Leaves of Grass*: Is Whitman's diction, in fact, innovative and what makes it so? How do Whitman's early poems compare to his later poems? What basis might be found for claims that Whitman is the great poet of America, women, the body, male homoeroticism, or democracy?

At the University of Nebraska-Lincoln's Center for Digital Research in the Humanities we have been experimenting with

a new way of visualizing phenomena in TEI corpora and have created Indigo, an experimental XSLT-based tool that queries TEI files and generates animated videos of the results. Using XPath and XQuery techniques, this tool makes it possible to ask specific or general questions of a corpus. The data are then output as scalable vector graphic (SVG) files that are converted to raster images and rendered in high definition H.264 video at 30 frames per second. At its core, Indigo is a program for performing scripted stop-motion animation, arranged in one or more scenes. What each scene contains is up to the user: it might include letters, numbers, shapes, colors, gradients, patterns, lines, paths, or imported raster images, each moving or not moving. The only requirement is that a scene must be modeled in XSLT, with SVG structures as the initial output. For the user wishing to visualize aspects of TEI text corpora, the news is good, for that format shares membership with XSLT and SVG in the XML ecosystem. Indigo provides a method for presenting, in fresh and unexpected ways, quantitative data relevant to scholarly questions in a way that is open-ended, making the user a co-creator with Whitman in the "meaning" of his texts.

Our experiment involves such activities as creating quantitative analyses of some of the linguistic characteristics of Whitman's poetic corpus, comparing them to those of some of his popular contemporaries, and then "presenting" the results as a video sequence. Such a procedure is admittedly outside the mainstream of critical methodology in the humanities, but it is entirely in keeping with Whitman's own theories of the proper relationships among authors, readers, and texts. "The process of reading," he said, "is not a half-sleep, but, in highest sense, an exercise, a gymnast's struggle; . . . the reader is to do something for himself, . . . must himself or herself construct indeed the poem, argument, history, metaphysical essay—the text furnishing the hints, the clue, the start or frame-work."¹

As Tanya Clement has recently observed, "sometimes the view facilitated by digital tools generates the same data human beings . . . could generate by hand, but more quickly," and sometimes "these vantage points are remarkably different . . . and provide us with a new perspective on texts."² And as Dana Solomon has written, "due in large part to its often powerful and aesthetically pleasing visual impact, relatively quick learning curve . . . and overall 'cool,' the practice of visualizing textual data has been widely adopted by the digital humanities."

In representing the literary work as an absorbing performance, one that comprises both "data" and "art," the method we are presenting is calculated to provoke responses in both informational and aesthetic registers. It is, in the terms of Jerome McGann and Lisa Samuels, an act of "interpretive deformance," whereby "we are brought to a critical position in which we can imagine things about the text that we didn't and perhaps couldn't otherwise know."³

References

1. **Whitman, Walt** (1892). *Democratic Vistas* in Complete Prose Works, (Philadelphia: David McKay), p. 257.
2. **Clement, T** (2013). *Text Analysis, Data Mining, and Visualizations in Literary Scholarship* in Literary Studies in the Digital Age: An Evolving Anthology (eds., Kenneth M. Price, Ray Siemens). Modern Language Association.
3. **Solomon, D.** (2013). *Building the Infrastructural Layer: Reading Data Visualization in the Digital Humanities*. MLA 2013 Conference Presentation. url: danaryansolomon.wordpress.com/2013/01/08/mla-2013-conference-presentation-from-sunday-162013/
4. **McGann, Jerome and Lisa Samuels**. *Deformance and Interpretation* url: www2.iath.virginia.edu/jjim2f/old/deform.html

CURIOS: Connecting and Empowering Community Heritage through Linked Data

Beel, David

University of Aberdeen, United Kingdom

Webster, Gemma

University of Aberdeen, United Kingdom

Mellish, Chris

University of Aberdeen, United Kingdom

Wallace, Claire

University of Aberdeen, United Kingdom

1. Moving Towards Community Digital Heritage

Rural areas are characterised by a strong identity of people with place. These identities draw on a repertoire of cultural norms, knowledge, histories, customs and practices which, taken together, construct unique place identities. This cultural distinctiveness is dynamic given traditional cultural practices are reproduced and others introduced as cultural systems evolve and adapt. Forms of cultural expression, such as storytelling, music and song, poetry and literature, dance and drama together with material objects, artefacts, sites and cultural spaces, are resources for interacting with the past and for experiencing the present. In the collection and transmission of these collections there has been a growing sense that the traditional methods for doing this are failing, Nora [1]. In order to address this problem, digital solutions have been sought but this has been a problematic process due to a number of variances. These include the constant changing of file types, software and codes of best practice, as well problems to do with cost and the sheer amounts of 'analogue' data to convert. Leading the way in this process have been national institutions but with the production of such local cultural repertoires, which as Flynn [2] suggests '*are the grassroots activities*' where '*control and ownership of the project is essential*' there has been a failure to consider the needs of community heritage groups in these processes. As such groups do not want to be subsumed into national archives, which they do not control, is not sensitive to their needs and is juxtaposed ideologically to the production of their own 'place history'. Following Creswell's [3] claim that such archives represent 'spaces of marginalized memory' CURIOS is therefore seeking a solution using open linked data in which a system can be developed that is attuned to the specificity of a local heritage but can also take advantage of already collected materials from elsewhere.

2. Case Study – Hebridean Connections

In the past 40 years around 22 'Comainn Eachdraidh'[1] (CE), have been established in the Outer Hebrides[2]. CE are community run groups that began in the 1970's with a very specific political and cultural purpose – to preserve the culture, history and language of the primarily Gaelic regions of Scotland. Such community heritage practices have been described as a 'messy' endeavour with a wide variety of different formal and informal practices [5]. The archives embrace different registers of social memory from tangible to intangible heritage, which have been collected and ordered in a variety of different ways. Different CE groups collect and order their archives in a variety of different ways: from the highly 'professional' to the more bespoke and sporadic. As the CE groups are voluntary community archives, they are rooted in local historical values, hence there is often little consistency between groups regarding cataloguing, archiving and content management.

Hebridean Connections (HC), which is a community managed, online historical resource was formed due to the driving force of a single member of a CE who saw the benefit of digitising and connecting the different historical catalogues [5]. The idea was proposed to the different CE and four groups were actively involved in a Heritage Lottery Fund (HLF) bid that funded the creation of the HC website[3]. The project website was launched in 2006, holding some 100,000 records relating to the genealogy, history, archaeology, and cultural traditions of the Outer Hebrides. Currently, the system allows users to search using keywords, selecting relevant images, or with a

map-based interface. Additionally, the website encourages contributions from its users and, therefore, has the potential to foster reciprocal knowledge exchange across geographical boundaries.

2.1 Sustainability

HC is one example of a community-built digital cultural heritage repository where their long-term future is unclear. Many issues with the current system have arisen since the initial grant, particularly surrounding funding and scalability. There is a real practical question about how a project of this kind can be maintained over time with the resources available to a small-dispersed community, especially as the initial system was developed by a private development company, using proprietary software. As the project developed, this situation raised the problem that any changes to the system required more financial investment in the software. For the small community heritage groups involved, this was not feasible, especially as the CE became aware of what was possible through digitisation and wanted to expand. The process of digitisation has created three primary issues for HC:

- How to expand the project remit without additional funding for developers?
- Scalability issues, how can more CE collections be integrated in a closed system
- Can empowering communities to control their own digital heritage improve long-term sustainability?

2.2 An Archive for the Future?

Motivated by the limitations of the current HC system, the CURIOS project's aim is to produce a sustainable system that allows a community of users to manage a digital archive of cultural heritage data, or 'cultural repository', releasing them from any specific proprietary software platform. To achieve this goal, CURIOS has made use of existing open source content management system (CMS) software and Semantic Web standards. The emergence of the Semantic Web [6] has led to several standard formats for representing and interchanging data [7, 8]. By making use of linked data, cultural repositories would have the potential for reuse and integration with further related data sources.

In recent years content management systems have gained popularity on the web by allowing users to build and publish web pages without requiring in-depth knowledge of the underlying web technologies. The CURIOS project has extended the web CMS approach to allow users to manage repositories of linked data. This Linked Data CMS approach makes use of existing CMS software to retain the usability and scalability of existing tools that are familiar to users, whilst allowing the users to exploit the benefits of linked data.

The Linked Data CMS approach has been implemented as a module for the popular open source Drupal CMS[1]. Building the next generation of Hebridean Connections on open source software and web standards has distinct advantages for future development and use of the system. The Drupal-based system can be maintained by its community of users and can be extended additional functionality developed by the Drupal open source community, e.g., to support blogging or e-commerce features. This community led maintenance allows for further future development of the cultural repositories as the archives develop.

3. Conclusion

Open linked data can help make local cultural repositories sustainable and collective. Linked data allows for collaboration, mutual authoring, distributed responsibilities through community projects and the utilisation of other community or national resources [4]. The CURIOS project is enabling local cultural heritage repositories to become a meaningful identity resource for an international community, who previously had no access to them. By falling outside of national institutional frameworks,

local people are the 'gatekeepers' of their own heritage and are selecting what to commemorate based on their own customs of remembering. This kind of digital archive can have, therefore, potentially significant social impacts which need to be better understood. The vision of Hebridean Connections is to expand the collections to incorporate those held by other Comainn Eachdraidh. Additionally, by making use of linked data, there is now the possibility to integrate further sources of data into HC from other historical societies or even national organisations.

4. Acknowledgments

We would like to thank Hebridean Connections and the Comainn Eachdraidh for their ongoing commitment to this research. This work is supported by the Rural Digital Economy Research Hub (EPSRC EP/G066051/1).

References

- Drupal is a popular open source web content management system: drupal.org .
- Comainn Eachdraidh is a Gaelic phrase meaning 'Historical Society'.
- The Outer Hebrides is a group of islands off the West coast of mainland Scotland.
- The Hebridean Connections website is hosted at www.hebrideanconnections.com .
- Nora, P.** (1996). *Realms of memory: rethinking the French past*. Volume 1: Conflicts and Divisions. Columbia: University Press.
- Flinn, A.** (2007). *Community Histories, Community Archives: Some Opportunities and Challenges 1* in *Journal of the Society of Archivists*. Volume 28, Issue 2.
- Creswell, T.** (2012) *Value, gleaning and the archive at Maxwell Street, Chicago* . *Transactions of the Institute of British Geographers*. Vol. 37 (1),1-13.
- Mellish, C., Wallace, C., Tait, E., Hunter, C., & MacLeod, M.** (2011). *Can Digital Technologies increase Engagement with Community History?*Digital Engagement 2011. de2011computing.dundee.ac.uk/wp-content/uploads/2011/10/Can-Digital-Technologies-increase-Engagement-with-Community-History.pdf
- Wallace, C., Tait, E., MacLeod, M., Mellish, C., & Hunter, C.** (2011). *Supporting Digital Humanities Creating Sustainable Digital Community Heritage Resources Using Linked Data*. In Supporting Digital Humanities: Answering theunaskable Conference. 17–18.
- Spector, A. Z.** (1989). *Achieving application requirements*. In Distributed Systems, S. Mullender, Ed. ACM Press Frontier Series. ACM, New York, NY, 19-33. DOI= doi.acm.org/10.1145/90417.90738 .
- Berners-Lee, T., Hendler, J., and Lassila. O.** (2001). *The Semantic Web*. Scientific American, 284(5), 34–43
- Hitzler, P., Krötzsch, M., Parsia, B., Patel-Schneider, P.F., and Rudolph, S.** (2009). (eds.) *OWL 2 Web Ontology Language: Primer*. W3C Recommendation, www.w3.org/TR/owl2-primer .

Unhappy? There's an App for That: Digital Happiness, Data Mining, and Networks of Well-Being

Belli, Jill

jbelli@citytech.cuny.edu
New York City College of Technology, CUNY

The content of well-being and the means for increasing it, at both individual and societal levels, are fundamentally utopian concerns. Consequently, positive psychology, the science of human flourishing, is an essentially utopian project: it intervenes in what it considers an unsatisfactory present

and attempts to create (and educate for) something better. Its explicitly activist and pragmatic agendas recently have been bolstered by the explosion of research into what I term “digital happiness.” In his 2011 book *Flourish: A Visionary New Understanding of Happiness and Well-Being*, leading positive psychologist Martin Seligman briefly outlines “positive computing” as using technological means and methods (such as data mining, social networking, and personalized apps) to “go beyond the slow progress in positive education to disseminate flourishing massively” (94).¹ Digital happiness proponents have taken up this call fervently, and positive psychologists, data analysts, coders, and policy makers regularly invoke the utopian possibilities of both technology and happiness as they collaborate for salvation on a grand scale: the H(app)athon Project hacks happiness in order to “save the world”²; video gamers and virtual reality developers (in Jane McGonigal’s words) create their products to “change the world”³; computer scientists offer their “Hedonometer” to measure, and therefore “improve or understand” well-being more completely;⁴ and the newly launched social network “Happier” will help “you feel freaking awesome.”⁵

Digital happiness initiatives are a fascinating site of inquiry because they combine the close reading and personal tracking aspects of the quantified self movement with the large scale data mining, aggregating, and visualization efforts associated with the digital humanities, “distant reading” (Moretti),⁶ and “network sense” (Mueller).⁷ Fueled by algorithms for happiness and subjecting qualitative phenomena to quantitative analysis (in ways reminiscent of Jeremy Bentham’s “felicific calculus” and the “mathematically infallible happiness” in Yevgeny Zamyatin’s classic dystopian novel, *We*⁸), these initiatives track and triangulate individual internal emotional states, networked virtual data and connections, and real social relations and policies. In this paper, I first provide an overview of these varied attempts to assess and maximize well-being using big data, sentiment analysis, crowdsourcing, social networking, the quantified self, and biometrics, positioning their rhetoric and ideology within the larger discourses of self-help, positive psychology and utopian studies. I then critically interrogate these projects’ utopian aspirations, analyzing their aims and methods for instantiating different ways of being and living, of creating both the happy individual and the good society in the image of (and from the) raw data of individuals’ emotions.

I ask what digital happiness methods, tools, applications, and findings teach us about what we should desire (and not desire), what we should value (and not value), what type of people we should be (and not be), and what type of actions we should take (and not take). Throughout the paper, I am in dialogue with Sara Ahmed’s notion of happiness as performative and normative in *The Promise of Happiness*(2010), and I highlight “not only what makes happiness good but how happiness participates in making things good” (13) and how “happiness shapes what coheres as a world” (2).⁹ In doing so, I not only critique digital happiness initiatives on an ideological level but also on technical and methodological levels. In particular, I interrogate happiness/well-being apps’ use of both active and passive data to fuel their algorithms; the methods of quantification, semantic analysis, and natural language processing in studies using social media to assess/analyze/improve happiness, well-being, and life satisfaction; how people interact with the technology that is tracking their happiness, and how these users often skew their responses in public, networked settings in order to present versions of their best selves to others. Ahmed has argued that happiness’s methods of self-reporting “both presumes the transparency of self-feeling (that we can say and know how we feel), as well as the unmotivated and uncomplicated nature of self-reporting. If happiness is already understood to be what you want to have, then to be asked how happy you are is not to be asked a neutral question” (5). In addition to this problem, already embedded within the positive psychology methodology, digital happiness assumes the transparency and translatability of language/texts and affect/emotions, and undertheorizes how the dynamics of a digital networked space change the way we communicate and connect with others. Digital happiness proponents advocate their work as contributing to the creation

of a more utopian future, and argue it is democratic because it is tracking raw data from the people themselves. But the questions of what raw data is assessed and who determines the metrics raise crucial questions about the type of vision these digital happiness experts put forth.

This paper also contextualizes my work on “digital happiness” within the larger discourses of both the self-help genre and positive psychology, and demonstrates how digital happiness showcases the competing tensions of individual improvement and social justice, apolitical progress and politically engaged action, and descriptive reporting and prescriptive advice in both. In doing so, I highlight positive psychology’s “discursive and political labor” (Yen 76).¹⁰ Particularly troubling is positive psychology’s conceptualization of its own politics and pedagogy, which teach us to be certain types of people in pursuit of the good life without consideration that its notion of the “good” is not morally universal but inextricably bound to the discipline’s ideological assumptions, cultural contexts (in particular, American individualism), and a particular interpretation of what is “positive,” valuable, and desirable.

I argue that, while digital happiness research’s use of big data and crowdsourcing partially tempers the rampant individualism that dominates positive psychology’s vision of the good life, its notions of the quantified self glorify the desirability of self-monitoring, normalcy, and discipline in the Foucauldian sense, which is a hallmark of much of the self-help genre (and positive psychology more generally). Self-help “render[s] social relations of power invisible and non-negotiable” and “counsels subjects to sculpt a meaningful life without addressing or questioning the horizon of social relations and the contexts of social power” (Rimke 65). Instead of serving as outlets for potential change, “[p]ractices of self-help are thus connected to the management and government of populations” (Rimke 72).¹¹ Similarly, self-help catches its users “a cycle of seeking individual solutions to problems that are social, economic, and political in origin” (McGee 177).¹² Therefore, its aggregated view of subjective well-being still sidesteps the important work of defining the ideological content/function of happiness and addressing its role in maintaining structural inequality.

This paper argues that positive psychology and, by extension, many digital happiness projects that are built on its values/methods, is inherently conservative, in the sense that it does not actively encourage radical possibility and transformation. While it is useful to identify and nurture the strengths that we already have, to reflect on experiences and create a positive meaning/communicate a positive message for them, and to derive satisfaction and pleasure from our activities/connections, these techniques run the risk of functioning remedially and driving us to become more complacent with “what is.” This limiting of possibility is one of the most troubling aspects of positive psychology’s work. As Levitas (2013) reminds us, the refusal to limit possibility is an essential part of the utopian project: “Utopia also entails refusal, the refusal to accept that what is given is enough. It embodies the refusal to accept that living beyond the present is delusional, the refusal to take a face value current judgements of the good or claims that there is no alternative” (17).¹³

I enact this utopian “refusal,” by arguing that positive psychology and digital happiness together create and endorse descriptions of and prescriptions for happiness and well-being that are quickly forming a unified front, a standardized, monolithic discourse that limits possibility. These fields matter, immensely, particularly because research on subjective well-being is being institutionalized prior to (or in some cases, in spite of) conversations about its assumptions, values, goals, and consequences. Therefore, this paper opens up a crucial conversation about what gets excluded in current discussions of well-being and digital happiness projects’ assessment and promotion of it.

References

1. **Seligman, Martin E. P.** (2011). *Flourish: A Visionary New Understanding of Happiness and Well-Being*. New York: Free Press. Print.
2. *The H(app)athon Project*. happathon.com/
3. **McGonigal, Jane** (2011). *Reality is Broken: Why Games Make Us Better and How They Can Change the World*. New York: Penguin.
4. *Hedonometer*. www.hedonometer.org/index.html
5. *Happier*. www.happier.com/
6. **Moretti, Franco** (2005). *Graphs, Maps, Trees: Abstract Models for a Literary History*. London: Verso. Print.
7. **Mueller, Derek** (2012). *Views from a Distance: A Nephological Model of the CCCC Chairs' Addresses, 1977-2011*. 'Kairos' 16.2 (Spring 2012). www.technorhetoric.net/16.2/topoi/mueller/
8. **Zamyatin, Yevgeny**. (1924). *We*. New York: EOS, 1999. Print.
9. **Ahmed, Sara** (2010). *The Promise of Happiness*. Durham, NC: Duke University Press. Print.
10. **Yen, Jeffrey**. (2010) *Authorizing Happiness: Rhetorical Demarcation of Science and Society in Historical Narratives of Positive Psychology*. Journal of Theoretical and Philosophical Psychology 30.2: 67-78.
11. **Rimke, Heidi Marie** (2000). *Governing Citizens Through Self-help Literature*. Cultural Studies 14.1: 61-78.
12. **McGee, Micki** (2007). *Self Help, Inc.: Makeover Culture in American Life*. New York: Oxford University Press. Print.
13. **Mueller, Derek** (2012). *Views from a Distance: A Nephological Model of the CCCC Chairs' Addresses, 1977-2011*. 'Kairos' 16.2 (Spring 2012). www.technorhetoric.net/16.2/topoi/mueller/

Where is my Other Half?

Ben-Shalom, Adiel
Friedberg Genizah Project

Choueka, Yaakov
Friedberg Genizah Project

Dershowitz, Nachum
nachum@cs.tau.ac.il
Tel Aviv University

Shweka, Roni
Friedberg Genizah Project

Wolf, Lior
Tel Aviv University

1. Introduction

One of the most challenging issues in the analysis of large collections of historical manuscripts or handwritten fragments is the "joining" of fragments, either piecing together torn parts of a mutilated folio – which may have disintegrated because of wear and tear over the ages – or reconnecting two or more folios originating from the same original manuscript, but which have since been separated and dispersed to diverse locations – perhaps on account of trading activities between institutions. A striking case in point is the Cairo Genizah, discovered towards the end of the nineteenth century in the attic of an old synagogue in Cairo, which contains more than 320,000 fragments (deriving from tens of thousands of individual documents, almost all in Hebrew characters – though not necessarily in the Hebrew language) spanning a thousand years of writing and copying. Various parts of the Genizah finds are currently located in more than sixty university collections and public libraries spread out on different continents all over the world. Until just recently, a Genizah researcher holding half a page in hand and seeking its other half, did not have any resources with which to achieve that goal beyond erudition, a few catalogs, a gifted memory, and a fair measure of luck.

In recent work ¹, we described an automated scheme for image analysis and processing that culminates in enabling the computer to compare two images and compute a similarity score based solely on the individual handwriting style. A series of benchmarks and tests convinced us of the reliability and utility of this metric. Moreover, many hundreds of "joins" of interest to humanities scholars have already been identified ². Here, we describe the design and implementation of a follow-up plan devised to integrate the matching scheme into a coherent and efficient system that can help scholars find the best candidates for potential joins for any given fragment. A system with somewhat similar goals for processing and joining fragments of frescoes is described in ³.

2. Joins and Jigsaws

The following steps have been implemented:

A – The basic idea is to match each of the Genizah fragments with one another so as to obtain a similarity score for each pair of fragments. We used a combination of local descriptors (SIFT) and learning techniques (OSS ⁴, SVM, and others). Out of an estimated total number of 320,000 fragments, about 230,000 fragments were available to us, represented by 450,000 digital images, with two images per fragment (recto and verso). For every fragment, a numerical signature vector was computed, encapsulating aspects of its writing style. With a specially designed software component that measures the readability of every fragment, we eliminated from this scenario most fragments with poor legibility, those that most likely would not contribute true joins but rather would deteriorate the effectiveness of the system. These included blank or almost-blank pages, illegible or very dark texts, minute fragments, etc. After eliminating these problematic items, we were left with a total of 158,000 fragments to be compared with one another. That gave a total of 12.4 billion pairs that needed to be measured for similarity, a huge number indeed. Some twenty different similarity scores were computed and stored for each pair. These were generated by using four different algorithms to represent the handwriting style of each document and by using different similarity measures between documents. The different similarity scores can be "stacked" together to achieve higher accuracy.

B – Twenty CPU's from the Computing Lab of the Blavatnik School of Computer Science at Tel Aviv University ran together continuously for 37 days (the equivalent of some 18,000 computing hours), and the task was accomplished. This computer run is probably one of the most intensive ever implemented in a digital humanities context, in terms of computing resources. Four terabytes of output were generated in the process.

C – An efficient and compressed database was built to preserve these results in a structure that is easy to manipulate within a reasonable on-line response time. For each fragment, the top 300 similar fragments were precomputed.

D – A simple program, *Propose Joins*, was then integrated in the operational software of the Genizah website, available at <http://www.jewishmanuscripts.org>. Any user can input an image number, and the system will respond immediately by giving a list of the best 100 candidates that might qualify as joins for the given fragment, sorted from most similar to least, accompanied by the actual images of these candidates. See Fig. 1. A user can then mark some images as worthy of further investigation as potential joins, passing them over to a second program, called "Jigsaw Puzzle", described below.

It is our assumption – backed up by Genizah researchers' expressed attitudes – that a competent user would not mind spending an hour or so examining these images, even if he or she does not end up finding any join in the set, since this is the only way to systematically look for such a join were there indeed any. Our experience shows, in fact, that if there is a join in the Genizah world for the given fragment, it will be found – almost always – in this set.

E – The *Jigsaw Puzzle* program displays the additional images designated by the user together with the original image on the screen, each image already restricted to the fragment's

physical contour, and, using the mouse and a few tabs, a user can magnify or reduce each of them, move any image, rotate it by any angle in any direction, "flip" the image over to display the verso (say) of the fragment instead of its recto, calibrate the images at their original proportions in order to check if the geometric features and the running text of the various pieces indeed fit neatly into a join. See Figs. 2-4. If a join is found, it may be incorporated in the website for all users to be made aware of it, with the identifying scholar's name inscribed as the join's composer.

F – To make the system even more user-friendly and intuitive to researchers, even if they are not completely at ease with computers, a large 42" touch-screen was installed in the Genizah lab, as a prototype, with an attached PC on which the website and the software were installed, all completely transparent to the user. See Fig. 5. Using a virtual keyboard, the user approaches the system by inputting an image number, receives back the images of the 100 best potential candidates, marks some of the relevant ones, passing them over to Jigsaw, where they can be easily manipulated – moved, rotated, flipped, calibrated – by just touching the screen with one's fingers, much like what one is used to doing nowadays with smartphones, and as naturally as one might arrange a jigsaw puzzle spread out on a table. See Fig. 6.

Discussion

We expect the overall scheme, with all the steps detailed above, to be of relevance in many other similar contexts, although, admittedly, the Genizah case is rather unusual in its scope and complexity. The join-matching tool is quite sophisticated and is constantly undergoing improvement. The jigsaw tool is relatively simple but has already proved very appealing to scholars. We are currently applying the join tool, more or less as-is, to other corpora, including the Dead Sea scrolls and papyri⁵ and Tibetan manuscripts and xylographs. Other potential applications include the 2,000,000 images of 70,000 pre-1900 Taiwanese deeds and court papers from the Taiwan History Digital Library⁶ and Yad Vashem's now publicly available Holocaust archives (www.yadvashem.org/yv/en/resources/index.asp). Furthermore, we hope to make the jigsaw tool more widely available. In addition, we have begun to use machine-learning tools based on the same signature vectors to help answer palaeographical questions for such corpora.^{7 8 9}

We are still left with the major problem of trying to reconstruct the original state of the entire Genizah collection, that is, to find, once and for all, the entire network of true joins in this collection, with a reasonable level of completeness and precision, and to do this efficiently and in a relatively short time. A crucial step in achieving this goal is to find effective methods for recognizing and eliminating large quantities of false joins and non-joins, even if that may be at the cost of losing a few correct ones. This is currently a topic of intense investigation, involving elements of graph theory, clustering techniques, data mining and related methodologies.

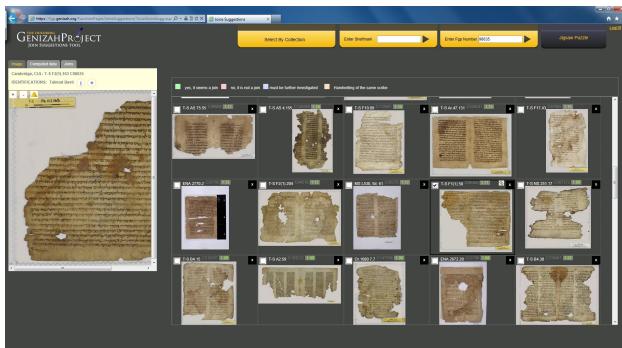


Fig. 1: After the user gave the system the i.d. number of the fragment being studied, the system responds by displaying that image on the left side, and 100 suggested joins, sorted by decreasing order of similarity, on the right.

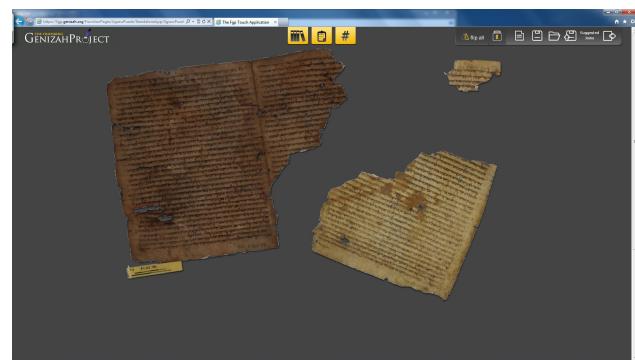


Fig. 2: The given fragment and two of the suggested joins.

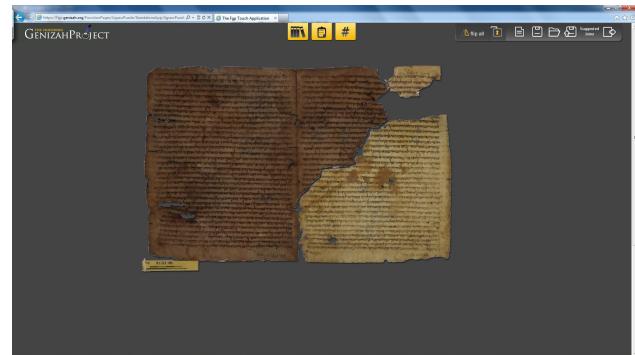


Fig. 3: Two fragments have been "glued together" by the Jigsaw program; the third is on its way.

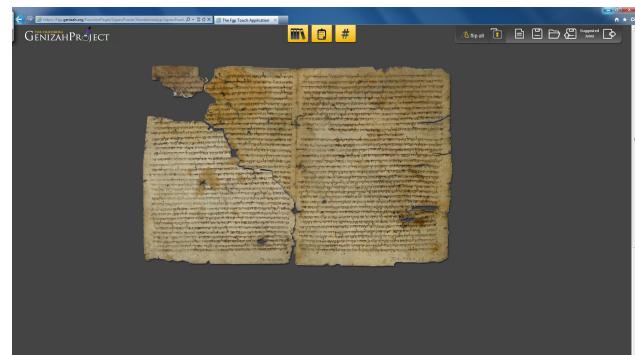


Fig. 4: The final join (verso).

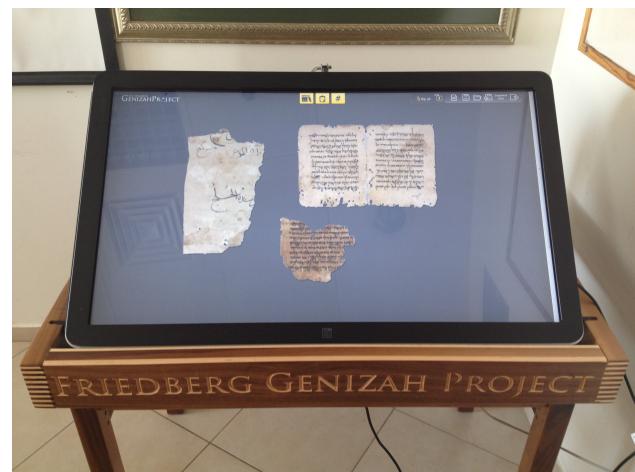


Fig. 5: The large touch screen mounted on a specially designed chassis, showing the fragment at hand together with 2 suggested joins.



Fig. 6: One of the authors (Y.C.) using the Jigsaw program to manipulate the suggested join.

References

1. Wolf, L., Littman, R., Mayer, N., German, T., Dershowitz, N., Shweka, R. and Choueka, Y. (2011). *Identifying Join Candidates in the Cairo Genizah*. International Journal of Computer Vision, 94(1): 118-135.
2. Shweka, R., Choueka, Y., Wolf, L. and Dershowitz, N. (2011). "Veqarev otam ehad el ehad": Zihuy ktav yad vetseruf qit'ei hagnizah beemtsa'ut mahshev (*Identifying Handwriting and Joining Genizah Fragments by Computer*), Ginzei Kedem, vol. 7, pp. 171-207. (In Hebrew.)
3. Brown, B. J., Toler-Franklin, C., Nehab, D., Burns, M., Dobkin, D. P., Vlachopoulos, A., Doumas, C., Rusinkiewicz, S. and Weyrich, T. (2008). *A System for High-Volume Acquisition and Matching of Fresco Fragments: Reassembling Theran Wall Paintings*. Proceedings SIGGRAPH 2008, ACM Trans. Graph., 27 (3).
4. Wolf, L., Hassner, T. and Taigman, Y. (2009). *The One-Shot Similarity Kernel*. IEEE International Conference on Computer Vision (ICCV), Kyoto, Japan, pp. 897-902, Sept. 2009.
5. Bearman, G. (2008). *Imaging the Dead Sea Scrolls for conservation purposes*. SPIE Newsroom, December 29, 2008.
6. Hsiang, J., Chen, S.-P. and Tu, H. C. (2009). *On Building a Full-Text Digital Library of Land Deeds of Taiwan*. Proceedings of Digital Humanities 2009, College Park, MD, June 2009, pp. 85-90.
7. Ben-Shalom, I. (2013). *Automatic Paleographic Grouping by Script Styles and Scribal Identity in Large Medieval Collections*. M.Sc. thesis, Tel Aviv University, Nov. 2013.
8. Dershowitz, N. and Wolf, L. (2013). *Automatic Scribal Analysis of Tibetan Writings*. Abstracts of the 13th Seminar of the International Association of Tibetan Studies, Ulaanbaatar, Mongolia, July 2013.
9. Wolf, L., Dershowitz, N., Potikha, L., German, T., Shweka, R. and Choueka, Y. (2011). *Automatic Paleographic Exploration of Genizah Manuscripts*. In: *Kodikologie und Paläographie im Digitalen Zeitalter 2 - Codicology and Palaeography in the Digital Age 2*, Fischer, F., Fritze, C. and Vogeler, G., eds., *Schriften des Instituts für Dokumentologie und Editorik*, vol. 3, Norderstedt: Books on Demand, Germany, pp. 157-179.

Marrying the Benefits of Print and Digital: Algorithmically Selecting Context for a Key Word

Benner, Drayton Callen

University of Chicago, United States of America

1. Introduction

Over the last few decades as more texts have been digitized, numerous software systems have arisen to display the texts and allow scholars to analyze them. These software systems have varied in their delivery (web-based, desktop software, mobile app, etc.) and their functionality, yet nearly all of them have included full-text search capabilities. Search is a central tool for scholars researching a corpus, and it is a task for which computers are perfectly suited. Despite the ubiquity of searching capabilities, there is no single method for displaying search results. When a user has requested to see a key word in its context, how much context should be presented to the user?

In choosing the context to present, there is no single solution that will always be best. At times, users will want to see detailed context requiring several lines. At other times, users will want to see as many search results as possible in a small visual space. Thus, providing multiple ways of viewing search results is desirable. When each search result is contained in a single line, perhaps the most attractive presentation currently in common use contains the key word in the middle, showing whatever context that fits on each side, a presentation style also found in some print concordances (Clarke, 1984; The Computer Bible, 1970-).

Occurrences 1-10:

ζω. υμᾶς μὲν οὖν ἐπιταῦν καὶ διαφερόντας ἀγαπῶ. ὅτι τῷ βίῳ μᾶλλον τῷ τῶν κρινομένων πιστεύετε. ἡ τοινότερον ἐπιπελέντεν ἐλάτικα. εἰς τοιούτον δὲ ἀγαπῶ συμβουλεύεις δὲ ἀμάλιστα κατεπείγει. νομίμη τῆς μὲν ὑπερβαλλοντας τῶν δὲ ἔργοντας οὐκ ἀγαπῶ ζῶν ἐπὶ τούτοις. αλλὰ οὐτα τοῖς γῆρασι ἔστι διστάσατον ἔχει οὐδέν. ἀλλὰ ἵνα γνῶν ὁ κόσμος ὃς ἀγαπῶ τὸν πατέρα, και καθὼς ἐντολὴν ἔδωκεν μοι ὁ πατέρας ἐκλεκτή κυρία και τοις τέκνοις αυτής, οὓς εἶναι ἀγαπῶ ἐν ἀληθείᾳ. και οὐκέτι μόνος ἀλλὰ και πάντες τοις κληματας τῆς Αχαΐας. δια τοις ὃς οὐκέτι μάρτυς, ο δεος οἰσεν. "Ο δε ποιῶ και ποτε τέρπων ψυχῶν μηδὲ εἰ περιστοτέρως μηδὲ ἀγαπῶ ἥστον ἀματόμα. "Εστο δέ, ἐγώ ουκ κατεβάρησα ὑμάτιον ἐποίησα τῆς νῦν οὐσίας· ἐγὼ δέ ἀγαπῶ ἐγώ μη ἐλάτιστα καταλύτοις τούτοισιν. ἀλλὰ βραχεῖ γένιον τοις κληματας πηγώς ἔστεσθαι. νῦν μὲν γάρ ἔγνωκ ἀγαπῶ ἦν γένιον τοις θεῖν δρμηθεῖς ἀνθρώπων μόνον

Fig. 1: . Results from a KWIC (Key Word in Context) search using Perseus under Philologic at perseus.uchicago.edu.

However, there is another method of displaying search results on a single line that is found in some print concordances that antedate digital tools. In this tradition, the context surrounding the key word is chosen manually so as to provide the reader as much information as possible about the key word's context.

strengthen	See also STRENGTHENED; STRENGTHENETH; STRENGTHENING.
De 3:28	and encourage him, and s' him: 553
Jdg 16:28	s' me, I pray thee, only this once, 2388
1Ki 20:22	Go, s' thyself, and mark, and see "
Ezr 6:22	to their hands in the work of the "
Ne 6: 9	therefore, O God, s' my hands.
Job 16: 5	I would s' you with my mouth, and 553
Ps 20: 2	sanctuary, and s' thee out of Zion, 5582
27: 14	and he shall s' thine heart: * 553
31: 24	courage, and he shall s' your heart,* "
41: 3	Lord will s' him upon the bed of *5582
68: 28 s'	O God, that which thou hast 5810
89: 21	mine arm also shall s' him. 553
119: 28 s'...me	according unto thy word. 6965
Isa 22: 21	robe, and s' him with thy girdle, 2388
30: 2	to s'...in the strength of Pharaoh, 5810
33: 23	they could not well s' their mast, 2388
35: 3	S' ye the weak hands, and confirm "
41: 10	I will s' thee; yea, I will help thee; 553
54: 2	thy cords, and s' thy stakes; 2388
Jer 23: 14	they s' also the hands of evildoers,
Eze 7: 13	s' himself in the iniquity of his life. "
16: 49	s' the hand of the poor and needy.
30: 24	s' the arms of the king of Babylon, "
25: 5	the arms of the king of Babylon, **"
34: 16	and will s' that which was sick:
Da 11: 1	I, stood to confirm and to s' him. 4581
Am 2: 14	and the strong shall not s' his force, 553
Zec 10: 6	And I will s' the house of Judah, 1396
12	And I will s' them in the Lord;
Lu 22: 32	art converted, s' thy brethren. *4741
1Pe 5: 10	perfect, establish, s', settle you. 4599
Re 3: 2	and s' the things which remain, *4741

Fig. 2: The entry "strengthen" in (Strong, 1890).

Unfortunately, this method requires a tremendous amount of manual effort; it has only been practical in concordances

of the Bible and other heavily-studied texts. The following concordances that antedate the maturation of the digital age take this approach: **Bible**: Cruden (1737); Young (1882); Strong (1890); Mandelkern (1896); Hazard (1922); Gant (1950); Lisowsky (1958); Even-Shoshan (1977); **Homer**: Prendergast (1875: Iliad); Dunbar (1880: Odyssey); **Shakespeare**: Clarke (1846). Since the full flowering of the digital age, it has been abandoned in most concordances and even in commercial Bible software, as shown in the following figures.

Result	Reference	Previous Context	Result	Next Context
NRSV	Ge 22:2	your son, your only son Isaac, whom you	love	, and go to the land of Moriah, and off to my master Abraham.
NRSV	Ge 24:12	ant me success today and show steadfast	love	to my master."
NRSV	Ge 24:14	I shall know that you have shown steadfast	love	to my master."
NRSV	Ge 24:27	aham, who has not forsaken his steadfast	love	and his faithfulness toward my master.
NRSV	Ge 24:67	ebekah, and she became his wife; and he	loved	her. So Isaac was comforted after his
NRSV	Ge 25:28	Isaac	loved	Esaу, because he was fond of game; b Jacob.
NRSV	Ge 25:28	because he was fond of game; but Rebekah	loved	.
NRSV	Ge 27:14	r prepared savory food, such as his father	loved	, and Rachel was graceful and beautiful.
NRSV	Ge 29:17	Leah's eyes were	lovely	Rachel; so he said, "I will serve you see
NRSV	Ge 29:18	Jacob	loved	he had for her.
NRSV	Ge 29:20	med to him but a few days because of the	love	Rachel more than Leah. He served Lat
NRSV	Ge 29:30	So Jacob went in to Rachel also, and he	loved	me."
NRSV	Ge 29:32	my affliction; surely now my husband will	love	and all the faithfulness that you have:
NRSV	Ge 32:10	not worthy of the least of all the steadfast	love	the girl, and spoke tenderly to her.
NRSV	Ge 34:3	was drawn to Dinah daughter of Jacob; he	loved	Joseph more than any other of his chi
NRSV	Ge 37:3	Now Israel	loved	him more than all his brothers, they h
NRSV	Ge 37:4	ut when his brothers saw that their father	loved	; he gave him favor in the sight of the
NRSV	Ge 39:21	as with Joseph and showed him steadfast	love	c

Fig. 3: Search results from Logos Bible Software (logos.com) on a PC.

Create KWIC/Collocation Table						
Copy Options		Version	Verse Range	Word	Left	Right
NRS	Gen 1:1 - 4Ma 18:24	love	5	5	2	Build Reload
Gen 22:2	only son isaac whom you	love	and go to the land			
Gen 24:12	success today and show steadfast	love	to my master abraham i			
Gen 24:14	I shall know that you have shown steadfast	love	to my master before he			
Gen 24:27	aham, who has not forsaken his steadfast	love	and his faithfulness toward my			
Gen 29:20	ebekah, and she became his wife; and he					
Gen 29:32	med to him but a few days because of the					
Gen 32:10	So Jacob went in to Rachel also, and he					
Gen 32:21	my affliction; surely now my husband will					
Gen 32:10	surely now my husband will	love	me she conceived again and			
Gen 32:10	least of all the steadfast	love	and all the faithfulness that			
Gen 39:21	joseph and showed him steadfast	love	he gave him favor in			
Exo 15:13	swallowed them in your steadfast	love	you led the people whom			
Exo 20:6	reject me but showing steadfast	love	to the thousandth generation of			
Exo 20:6	thousandth generation of those who	love	me and keep my commandments			
Exo 21:5	if the slave declares i	love	my master my wife and			
Exo 34:6	anger and abounding in steadfast	love	and faithfulness keeping steadfast love			
Exo 34:7	love and faithfulness keeping steadfast	love	for the thousandth generation forgiving			
Lev 19:18	your people but you shall	love	your neighbor as yourself i			
Lev 19:34	citizen among you you shall	love	the alien as yourself for			
Num 14:18	anger and abounding in steadfast	love	forgiving iniquity and transgression but			

Fig. 4: . Search results from BibleWorks (BibleWorks.com) on a PC.

Search: NRSV		(Found: 540)	New
love			
Exodus			
15:13	13 "In your steadfast love you led the people whom you		
	redeemed;		
20:6	but showing steadfast love to the thousandth		
	generation ^b of those who love me and keep my		
21:5	5 But if the slave declares, "I love my master, my wife,		
	and my children; I will not go out a free person,"		
	6 then		

Fig. 5: Search results from Olive Tree Bible Software (OliveTree.com) on a Samsung Galaxy S3 smartphone. As a disclaimer, I wrote the search engine—but not the code to display the search results—for Olive Tree Bible Software as an independent contractor.

There have been some print concordances in the digital age for which the single-line context has been produced algorithmically, at least in part (e.g. Ellison, 1957; Spevack, 1968-1975; Goodrick and Kohlenberger, 1990; Kohlenberger, 1991; Dixon and Dawson, 1992; Mounce, 2012). Where algorithmic details have been published in part (Soule, 1956; Dixon, 1974; Dawson, 1977; Burton, 1982), they have often been primarily dependent on punctuation and/or manual annotation as a pre-processing step. While pioneering in their day, computing resources are more plentiful today, and the field of natural language processing has advanced greatly.

I propose an algorithm that seeks to mimic a human reader's choice of context for a search term. The goal is to produce the most relevant context for a key word on a line that is of arbitrary width using an arbitrary font. This provides the benefits

of traditional print concordances without the tens or hundreds of person-years required to produce them for even a single line width and font size.

2. Algorithm

2.1 Preprocessing

The text must, of course, be available in electronic form, but a syntactic parsing is also necessary. There are some electronic texts that have been parsed syntactically by hand (e.g. Andersen and Forbes, 2012), but the recent development of general-purpose parsers has made this work possible on a broader scale. As these parsers are developed for more languages and dialects and as they improve, the approach outlined here will become more and more useful. For this work, I generated phrase structure trees and dependency trees using StanfordCoreNLP (version 1.3.5, nlp.stanford.edu/software/index.shtml) on three texts (cf. Toutanova et al., 2003; de Marneffe et al., 2006). In keeping with the traditional use of concordances with Bible translations, I chose two Bible translations along with one novel: the *King James Version (KJV)* of the Bible (1769 text edition), the *English Standard Version (ESV)* of the Bible (2011 text edition, Old Testament/Hebrew Bible portion only), and Henry James' novel *What Maisie Knew (Maisie)*. A small amount of preprocessing was done before and after StanfordCoreNLP's parsing, both to fix some repetitive errors in StanfordCoreNLP's analysis and also to remove, and then reinstate, the main archaisms in the KJV.

2.2 Algorithm

In order to develop an algorithm for mimicking a human's choice of context, I developed training data by randomly choosing key words from the *ESV* and line lengths, ranging from what might fit legibly on a typical smartphone to a line three times as long. I then displayed all possible contexts that fit on the line but make maximum use of the space on the line. That is, no more words could fit on either side. In addition, sensible rules concerning which types of punctuation were appropriate at the beginning or end were employed (e.g. a possible context could not begin with a comma or end with an open quotation mark), and possible contexts could not cross verse boundaries. In the rare case that there was only one option, that key word was discarded. A user selected his preferred context for 500 such key words, occasionally choosing two or three different contexts if they seemed equally desirable¹. After analyzing his choices, I produced the following metric $w(k,n)$ to give a value (weight) to each nearby word n for the key word k :

$$\text{Let } w(k,n) = \begin{cases} 4^{v(k,n)}, & v(k,n) \geq 0 \\ -4^{-v(k,n)}, & v(k,n) < 0 \end{cases}, \text{ where}$$

$$\begin{aligned} v(k,n) &= c_p * p(k,n) + c_f * f(n) + s(k,n) \\ p(k,n) &= \text{the number of punctuation marks between } k \text{ and } n; \\ &\quad (\text{contiguous punctuation marks count as 1}) \end{aligned}$$

$$f(n) = \begin{cases} 1, & n \text{ is a content word} \\ 0, & n \text{ is a function word} \end{cases}$$

$$s(k,n) = 10 * 0.93(c_{pk}*d_{pk} + c_{pn}*d_{pn} + c_{dk}*d_{dk} + c_{dn}*d_{dn})$$

Let a_p be the nearest common ancestor of k and n in the phrase structure tree and a_d be the nearest common ancestor of k and n in the dependency tree. Then d_{pk} is the distance between a_p and k in the phrase structure tree, d_{pn} is the distance between a_p and n in the phrase structure tree, d_{dk} is the distance between a_d and k in the dependency tree, and d_{dn} is the distance between a_d and n in the dependency tree.

Each possible context is evaluated as the sum of $w(k,n)$ for each n in the possible context; the context with the highest value is chosen. If multiple possible contexts have identical

values, any can be chosen; I picked the one that had the most context before the key word.

The constants were optimized to the following values using a Monte Carlo particle filter on the training data:

$$c_f = 1.5; c_p = -3.37; c_{pk} = 0.175; c_{pn} = 0.2; c_{dk} = 1; c_{dn} = 1.4.$$

These constants reveal that the dependency tree was more important than the phrase structure tree.

3. Results

In addition to the above-mentioned training set, test sets were then generated from the *ESV* and *Maisie* with 100 key words each, and four human annotators made selections for each. The results are listed in Table 1. Since human annotators occasionally selected two or three contexts as equally good, a match for a given key word is calculated as:

$$\frac{|C_1 \cap C_2|}{\min(|C_1|, |C_2|)}, \text{ where } C_1 \text{ and } C_2 \text{ are the sets of contexts selected by the annotators.}$$

	<i>ESV</i> training set	<i>ESV</i> test set	<i>Maisie</i> test set
Algorithm matches user selection	67.8%	62.5%	47.8%
Expected algorithm matches if selections were random from a uniform distribution	27.4%	25.5%	21.9%
Inter-annotator agreement	N/A	65.8%	53.0%
Expected inter-annotator agreement if selections were random from a uniform distribution	N/A	27.0%	23.5%

These results indicate that on average, the algorithm matches a given human annotator slightly less often than another human annotator does. Assuming that human intuition presents the gold standard for this task, this means that the algorithm is doing only slightly worse than humans at picking the best context for the key word.

Some screenshots of algorithmically-generated key words in context are shown below.

Reference	KWIC
Exodus 4:14	he said, "Is there not A, your brother,
Exodus 4:27	The LORD said to A, "Go into the
Exodus 4:28	And Moses told A all the words of the
Exodus 4:29	A spoke all the words which the LORD
Exodus 5:1	Moses and A went and said to Pharaoh
Exodus 5:4	said to them, "Moses and A, why do
Exodus 5:20	They met Moses and A, who were
Exodus 6:13	spoke to Moses and A and gave them a
Exodus 6:13	sister, and she bore him A and Moses.
Exodus 6:13	A took as his wife Elisheba; the
Exodus 6:26	These are the A and Moses to whom
Exodus 7:1	you belong." They said to him, "A,
Exodus 7:1	your brother A shall be your prophet.
Exodus 7:2	your brother A shall tell Pharaoh to let
Exodus 7:6	Moses and A did so; they did just as
Exodus 7:7	you said, and A eighty-three years old
Exodus 7:8	Then the LORD said to Moses and A,
Exodus 7:9	then you shall say to A, 'Take your
Exodus 7:9	Moses and A went to Pharaoh and did
Exodus 4:14	not A Levite your brother?
Exodus 4:27	And the LORD said to A, Go
Exodus 4:28	Moses told A all the words of the
Exodus 4:29	A spoke all the words which the LORD
Exodus 5:1	Moses and A went and gathered
Exodus 4:30	And A spake all the words which
Exodus 5:1	afterward Moses and A went in,
Exodus 5:4	do you, Moses and A, let the
Exodus 5:20	And they met Moses and A, who
Exodus 6:13	spake unto Moses and unto A,
Exodus 6:20	and she bare him A and Moses;
Exodus 6:23	And A took him Elisheba;
Exodus 6:25	And Eleazar's son took him
Exodus 6:26	These are that A and Moses, to
Exodus 6:27	these are that Moses and A.
Exodus 7:1	A and your brother shall be your
Exodus 7:2	A your brother shall speak unto
Exodus 7:6	Moses and A did as the LORD
Exodus 7:7	A fourscore and three years old,

Fig. 6: KWIC search for "Aaron" in *ESV*, *KJV*; "*Maisie*" in *Maisie* (from left to right).

Reference	KWIC	Surface form
Genesis 8:15	ark, you and your wife, and your sons,	ark, you and your wife, and your sons,
Genesis 8:15	bring a remnant in the land,	bring a remnant in the land,
Genesis 30:26	Sheba, And he knew her spinney	sheba, and he knew her spinney
Exodus 16:23	make what you will have and what will	make what you will have and what will
Exodus 19:12	I can not go up into the mountain; take	i can not go up into the mountain; take
Exodus 22:24	Any animal that has its testicles bruised;	any animal that has its testicles bruised;
Leviticus 27:30	All the land, whether of the	all the land, whether of the
Numbers 23:23	God brings him of Egypt and so for out	god brings him of egypt and so for out
Numbers 31:27	two parts it is warlike who were between	two parts it is warlike who were between
Deuteronomy 6:2	I command you, whole	i command you, whole
Deuteronomy 26:5	shall come a yoke and overtake you, if you	shall come a yoke and overtake you, if you
Joshua 22:21	in the people of Peleth a the people and	in the people of peleth a the people and
Judges 20:20	the men went out to fight of	the men went out to fight of
Judges 20:28	They met Moses and A, who were	they met moses and a, who were
Joshua 2:1	He said to him, "I am a servant of the	he said to him, i am a servant of the
1 Samuel 3:1	The song of God had a yet gone out,	the song of god had a yet gone out,
1 Samuel 3:23	that he may be anointed of the king's	that he may be anointed of the king's
1 Samuel 38:25	he	he
Genesis 1:15	the heaven I give side upon the earth;	the heaven i give side upon the earth;
Genesis 1:25	And every thing of the gold colour it was	and every thing of the gold colour it was
Genesis 2:1	the LORD God had called it	the lord god had called it
Genesis 5:31	were seven hundred and seven years	were seven hundred and seven years
Genesis 8:5	decreased continually until it reached the	decreased continually until it reached the
Genesis 10:1	land of Canaan;	land of canaan;
Genesis 10:25	And he said, "C is for Canaan, a servant of Cursed	and he said, c is for canaan, a servant of cursed
Genesis 10:19	He is a mighty hunter before the	he is a mighty hunter before the
Genesis 13:14	the land of Canaan; and so for out	the land of canaan; and so for out
Genesis 18:31	I will not destroy it for twenty years,	i will not destroy it for twenty years,
Genesis 19:19	And I pressed upon the man, even they	and i pressed upon the man, even they
Genesis 22:20	things, that it was I Abraham, saying	things, that it was i abraham, saying
Genesis 27:27	and bless you is the LORD before me before	and bless you is the lord before me before
Genesis 28:13	and he rolled the stone from the well;	and he rolled the stone from the well;
Genesis 30:26	you know my service which I have done	you know my service which i have done
Genesis 35:18	give you power w God, and the land with	give you power w god, and the land with
Genesis 35:18	the house and the land, and the house and the land,	the house and the land, and the house and the land,
Genesis 36:43	according to their habitations in the land to	according to their habitations in the land to
Genesis 41:1	of two full years, that P'dan-	of two full years, that p'dan-
1 Samuel 4:11	Pharao	pharao

Reference	KWIC	Surface form
Genesis 8:15	the heaven I give side upon the earth;	the heaven i give side upon the earth;
Genesis 1:25	And every thing of the gold colour it was	and every thing of the gold colour it was
Genesis 2:1	the LORD God had called it	the lord god had called it
Genesis 5:31	were seven hundred and seven years	were seven hundred and seven years
Genesis 8:5	decreased continually until it reached the	decreased continually until it reached the
Genesis 10:1	land of Canaan;	land of canaan;
Genesis 10:25	And he said, "C is for Canaan, a servant of Cursed	and he said, c is for canaan, a servant of cursed
Genesis 10:19	He is a mighty hunter before the	he is a mighty hunter before the
Genesis 13:14	the land of Canaan; and so for out	the land of canaan; and so for out
Genesis 18:31	I will not destroy it for twenty years,	i will not destroy it for twenty years,
Genesis 19:19	And I pressed upon the man, even they	and i pressed upon the man, even they
Genesis 22:20	things, that it was I Abraham, saying	things, that it was i abraham, saying
Genesis 27:27	and bless you is the LORD before me before	and bless you is the lord before me before
Genesis 28:13	and he rolled the stone from the well;	and he rolled the stone from the well;
Genesis 30:26	you know my service which I have done	you know my service which i have done
Genesis 35:18	give you power w God, and the land with	give you power w god, and the land with
Genesis 35:18	the house and the land, and the house and the land,	the house and the land, and the house and the land,
Genesis 36:43	according to their habitations in the land to	according to their habitations in the land to
Genesis 41:1	of two full years, that P'dan-	of two full years, that p'dan-
1 Samuel 4:11	Pharao	pharao

Fig. 7: Randomly Selected Key Words from *ESV*, *KJV* and *Maisie* (from left to right).

4. Conclusion

Searching for key words is one of the core functions of text analysis software. The work presented here holds promise as a way of improving the way in which search results are displayed by automating a time-consuming manual technique traditionally used in print concordances. In addition, future work could deal with more complex displays, including possibly not using all the space available, possibly using ellipses, and dealing with displaying results of searches involving multiple key words.

Notes

1. I would like to thank James Covington for his annotation of the training set and both test sets, Rodelle Williams and D. Chris Benner for their annotation of both test sets, Humphrey H. Hardy for his annotation of the *ESV* test set, and Samuel L. Boyd for his annotation of the *Maisie* test set.

References

- Andersen, F. I. & Forbes, A. D. (2012). Biblical Hebrew Grammar Visualized. Winona Lake: Eisenbrauns.
- Burton, D. M. (1982). Automated Concordances and Word-indexes: Machine Decisions and Editorial Revisions. Computers and the Humanities 16, 195-218.
- Clarke, E. G. (1984). Targum Pseudo-Jonathan of the Pentateuch: Text and Concordance. Hoboken: Ktav.
- Clarke, M. C. (1846). The Complete Concordance to Shakespeare: Being a Verbal Index to all the Passages in the Dramatic Works of the Poet. New York: Wiley and Putnam.
- Cruden, A. (1737). A Complete Concordance to the Old and New Testament; or a Dictionary and Alphabetical Index to the Bible with a Concordance to the Apocrypha, and a Compendium of the Holy Scriptures. London: Frederick Warne & Co.
- Dawson, J. L. (1977). Textual Bracketing. ALLC Bulletin 5, 148-157.
- de Marneffe, M.-C., MacCartney, B. & Manning, C. D. (2006). Generating Typed Dependency Parses from Phrase Structure Parses. Language Resources and Evaluation Conference. Genoa, Italy.
- Dixon, J. E. G. (1974). A Prose Concordance: Rabelais. ALLC Bulletin 2, 47-54.
- Dixon, J. E. G. & Dawson, J. L. (1992). Concordance des Œuvres de François Rabelais. Genève: Droz.
- Dunbar, H. (1880). A Complete Concordance to the Odyssey and Hymns of Homer. To which is added A Concordance to the Parallel Passages in the Iliad, Odyssey and Hymns. Oxford: Clarendon.
- Ellison, J. W. (1957). Nelson's Complete Concordance of the Revised Standard Version of the Bible. New York: Nelson.
- Even-Shoshan, A. (1977). Konkordantsyah hadashah le-Torah, Nev'im, u-Khetuvim: botsar leshon ha-Mikra - Ivrit ya-Aramit: shorashim, milim, shemot peratim, tserufim ve-nirdafim. Jerusalem: Kiryat sefer.
- Gant, W. J. (1950). Concordance of the Bible in the Moffatt Translation. London: Hodder and Stoughton.
- Goodrick, E. W. & Kohlenberger, J. R., III (1990). The NIV Exhaustive Concordance. Grand Rapids: Zondervan.

- Hazard, M. C.** (1922). A Complete Concordance to the American Standard Version of the Holy Bible. New York: Nelson.
- Kohlenberger, J. R., III** (1991). The NRSV Concordance Unabridged: Including the Apocryphal/Deuterocanonical Books. Grand Rapids: Zondervan.
- Lisowsky, G.** (1958). Konkordanz zum hebräischen Alten Testament, nach dem von Paul Kahle in der Biblia Hebraica edidit R. Kittel besorgten masoretischen Text. Stuttgart: Privileg Württ. Bibelanstalt.
- Mandelkern, S.** (1896). Veteris Testimenti concordantiae hebraicæ atque chaldaicæ, quibus continentur cuncta quae in prioribus concordantiis reperiuntur vocabula, lacunis omnibus expletis, emendatis cuiusquemodi vitiis, locis ubique denuo excerptis atque in meliorem formam redactis, vocalibus interdum adscriptis, particulae omnes adhuc nondum collatae, pronomina omnia hic primum congesta atque enarrata, nomina propria omnia separatim commemorata. Lipsiae: Veit et comp.
- Mounce, W. D.** (2012). ESV Comprehensive Concordance of the Bible. Wheaton: Crossway.
- Prendergast, G. L.** (1875). A Complete Concordance to the Iliad of Homer. London: Longmans, Green & Co.
- Soule, G.** (1956). Machine that Indexed the Bible. Popular Science 169, 173-175, 242, 246.
- Spevack, M.** (1968-1975). A Complete and Systematic Concordance to the Works of Shakespeare. Hildesheim: Georg Olms.
- Strong, J.** (1890). The Exhaustive Concordance of the Bible: Showing every Word of the Text of the Common English Version of the Canonical Books, and every Occurrence of each Word in Regular Order: together with A Comparative Concordance of the Authorized and Revised Versions, Including the American Variations: Also Brief Dictionaries of the Hebrew and Greek Words of the Original, with References to the English Words. Cincinnati: Jennings & Graham.

- The Computer Bible.** (1970-). Missoula: Scholars Press.
- Toutanova, K., Klein, D., Manning, C. D. & Singer, Y.** (2003). Feature-rich Part-of-speech Tagging with a Cyclic Dependency Network. Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1. Edmonton, Canada: Association for Computational Linguistics.
- Young, R.** (1882). Analytical Concordance to the Bible on an Entirely New Plan: Containing every word in Alphabetical Order, Arranged under its Hebrew or Greek Original, with the Literal Meaning of each, and its Pronunciation; Exhibiting about Three Hundred and Eleven Thousand References, Marking 30,000 Various Readings in the New Testament, with the Latest Information on Biblical Geography and Antiquities, etc. etc. etc. Philadelphia: Lippincott & Co.

Riorganizzare SignWriting per favorire la ricerca linguistica sulle Lingue dei Segni

Bianchini, Claudia S.
 claudia.savina.bianchini@univ-poitiers.fr
 EA3816 FoReLL - Université de Poitiers

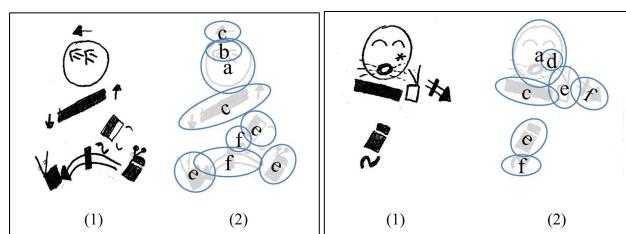
Borgia, Fabrizio
 fabrizio.borgia@uniroma1.it
 "Sapienza" Università di Roma

De Marsico, Maria
 demarsico@di.uniroma1.it
 "Sapienza" Università di Roma

Le Lingue dei Segni (LS) sono lingue utilizzate dalla maggior parte dei sordi del mondo per comunicare tra di loro; si tratta di lingue storico-naturali caratterizzate dall'uso del canale visuo-gestuale. La loro rappresentazione grafica è una sfida per la ricerca linguistica in quanto le LS non hanno ancora sviluppato un sistema di scrittura condiviso dalla comunità e,

al contempo, non possono essere rappresentate adattando un sistema di scrittura nato per le lingue vocali (come l'alfabeto latino o anche l'alfabeto fonetico internazionale) a causa della differenza di canale comunicativo e della conseguente multi-linearità della lingua. Tale assenza provoca difficoltà nei compiti di notazione (ossia rappresentazione delle forme significanti), annotazione (ossia rappresentazione dei fenomeni linguistici) e successiva analisi di produzioni segnate.

Attualmente, la maggior parte delle analisi delle LS si basa sull'uso delle cosiddette "glosse", etichette verbali nella lingua vocale del ricercatore che permettono - secondo il loro scopo e il loro grado di complessità - di annotare il significato del segno, le caratteristiche della sua forma significante o anche altre informazioni che lo riguardano. Sebbene possa trattarsi di etichette anche molto precise, rimane difficile – se non impossibile – ricostruire la forma significante delle produzioni segnate attraverso queste etichette: si tratta quindi di soluzioni che permettono l'annotazione ma non la notazione delle LS. La soluzione usata, che aggira il problema della rappresentazione grafica della lingua, è quella di associare temporalmente a tali etichette il video originale, grazie all'uso di software di annotazione linguistica come ELAN¹ o ANVIL².



Segni rappresentati con SW (1) e specifica degli elementi descritti (2):
 (a) espressione facciale, (b) direzione dello sguardo, (c) movimento e posizione della testa e del corpo, (d) contatto delle mani, (e) configurazione delle mani, (f) movimento e posizione delle mani e delle braccia

Tra i non molti sistemi esistenti per rappresentare graficamente le LS vi è SignWriting (SW) di V. Sutton³, nato per raffigurare le posizioni ed i movimenti del corpo (in particolare nella danza) e poi applicato alle LS. E' stato mostrato⁴ che esso rappresenta una valida soluzione in quanto consente sia la scrittura che la trascrizione di produzioni segnate, mantenendo le caratteristiche delle LS e permettendo una buona leggibilità dei testi così concepiti. Tale strumento può quindi essere utilizzato sia dai ricercatori a fini scientifici che dalla comunità sorda in generale.

Dal punto di vista del sistema grafico, SW è organizzato in 37 811 glifi (unità grafiche minime) che vengono disposti in una vignetta (spazio bidimensionale che rappresenta analogicamente lo spazio segnico tridimensionale) per comporre un segno [Fig. 1]. Ogni glifo rappresenta una componente manuale (configurazione, movimento, coordinazione, etc.) o non-manuale (espressione del viso, movimento del corpo, etc.) del segno. E' così possibile rappresentare sia le unità lessematiche⁵ che le strutture di grande iconicità⁶ sia segni isolati che segni inseriti in un discorso.

L'ostacolo maggiore incontrato da SW è la sua scarsa diffusione, dovuta in gran parte al numero esorbitante di glifi che lo compongono: tale quantità sarebbe, infatti, sufficiente a scoraggiare anche l'apprendente più volenteroso. E' stato tuttavia dimostrato^[4] che, una volta passata la diffidenza iniziale, è possibile – per soggetti segnanti - cominciare ad utilizzare il sistema in poche ore, grazie ad una forte regolarità di questi glifi. In ogni caso, il numero considerevole di glifi pone problemi anche alla digitalizzazione delle produzioni in SW ed ad un loro riconoscimento automatico, il che si riflette negativamente sulla possibilità concreta di usare SW a scopo di ricerca, per esempio per analizzare il legame tra la forma di un segno ed i fenomeni linguistici.

Il lavoro qui proposto presenta una radicale riorganizzazione di SW (d'ora in poi chiamata SW^{clt¹}) per raggruppare i glifi secondo le loro caratteristiche comuni e per mettere in evidenza quelle regole implicite che permettono un suo facile

apprendimento nonostante la quantità di glifi esistenti. A tale scopo si è provveduto a raccogliere i glifi sulla base della loro funzione (configurazione, movimento della mano, direzione dello sguardo...) per poi andare a individuare le caratteristiche che consentono di passare da un glifo prototípico (GP) ad un glifo variazione (GV). Ad esempio, per le configurazioni, sono stati individuati 242 GP e 4 regole di variazione con relativi set di possibilità (mano x 2 possibilità, piano x 2, lato x 3 e rotazione x 8) [Fig. 2]. I GP sono anche stati classificati in modo da evidenziarne le somiglianze, e così, imparando le 4 regole e circa una trentina dei 242 GP, è possibile sfruttare 23 232 GV ossia oltre la metà dei glifi esistenti. La facilità di apprendimento di SW dopo la riorganizzazione con SW^{rec1} è stata preliminarmente testata su un gruppo di studenti udenti di LS: i risultati sono incoraggianti in quanto, dopo appena 6 ore di formazione, i soggetti erano in grado di orientarsi facilmente all'interno di SW^{rec1} (cosa non possibile con la vecchia classificazione) e di comporre e leggere brevi testi scritti in SW.

G _P ↓	Mano destra			Mano sinistra		
	Palmo	Taglio	Dorso	Palmo	Taglio	Dorso
Piano verticale						
Piano orizzontale						

Ogni glifo prototípico che identifica una configurazione può, per trasformazione, dar vita a 96 glifi variazione basati sulla mano utilizzata e il suo orientamento

Oltre alle implicazioni nell'apprendimento di SW, SW^{rec1} mira anche a favorire la sua informatizzazione, sia per permetterne un più facile utilizzo da parte degli utenti sordi e udenti, che per renderlo uno strumento utilizzabile nella ricerca linguistica. A tale scopo è nata la collaborazione tra gli autori - un linguista e due informatici - con l'idea di cercare di risolvere, un passo alla volta, le difficoltà legate alla digitalizzazione del sistema grafico.

In primo luogo, è stato creato SWift (SignWriting Improved Fast Transcriber⁷), un editor di testi in SW, che permette attraverso una struttura ed un'interfaccia "deaf-centered" (ossia interamente pensata e realizzata in collaborazione con gli utenti sordi) di produrre testi in SW. In SWift l'utilizzo di SW^{rec1} ha permesso di rendere più rapido, efficace e "user friendly" lo strumento di ricerca dei glifi che verranno usati dall'utente per comporre il proprio testo.

Nel corso delle loro ricerche su SW, gli autori hanno notato la tendenza degli utenti osservati a comporre prima a mano e poi a ricopiare al computer i testi: ciò ha portato alla necessità di implementare applicazioni di OGR (Optical Glyph Recognition), con caratteristiche diverse dal classico OCR (Optical Character Recognition) in quanto più vicine al riconoscimento di scritture di tipo ideografico. In effetti, OGR deve prendere in considerazione un numero molto elevato di glifi, organizzati in uno spazio grafico non lineare e diposti secondo regole di composizione non rigide. Una volta concluso il suo sviluppo, OGR effettuerà il riconoscimento basandosi esclusivamente sulle caratteristiche geometriche e topologiche dei glifi; si è pertanto resa necessaria - in fase di progettazione - una seconda fase di riclassificazione di SW (SW^{rec2}) basata su queste caratteristiche.

Le applicazioni menzionate, basate su SW^{rec1} e SW^{rec2}, dovrebbero permettere una facile digitalizzazione di documenti prodotti in SW, il che favorirà la diffusione del sistema grafico presso la comunità sorda. Tuttavia, ciò che preme ai linguisti che usano SW è che esso possa finalmente diventare uno strumento "utilizzabile" per fare analisi sulle forme della LS. La poca integrazione, fino ad oggi, tra SW ed i software di annotazione linguistica più comunemente usati implica, infatti, una seria difficoltà per sfruttare le potenzialità di SW

come sistema di notazione delle LS. Integrando SW con ELAN, sarebbe possibile mettere in evidenza le relazioni tra significante e significato, le correlazioni tra forme del segno e strutture linguistiche, etc. ossia tra la notazione e l'annotazione delle LS.

Oltre alla sfida dell'integrazione informatica, si pone un duplice problema legato alla natura stessa di SW: visto il numero elevato di glifi e la scarsa standardizzazione di utilizzo del sistema grafico, è difficile realizzare trascrizioni che permettano poi un effettivo confronto sulle forme. Tale problema è però superabile grazie alla SW^{rec1} che - mettendo in evidenza le caratteristiche intrinseche ai singoli glifi (caratteristiche trasversali a diversi GP) - permette di collegare tra di loro realizzazioni grafiche diverse di forme segnificate simili, o anche caratteristiche identiche in forme segnificate diverse.

SW^{rec1} ed SW^{rec2} si pongono quindi così come primi passi per un'informatizzazione di SW che dovrebbe permettere lo sviluppo e la diffusione del sistema sia all'interno della comunità sorda che nell'ambito della ricerca linguistica sulla LS.

References

1. **ELAN Team**. 2002. *ELAN - Electronic Linguistic ANnotator*. Nijmegen (Netherlands), Max Planck Institute of Psycholinguistics. www.lat-mpi.eu/tools/elan/ [software, disponibile on line].
2. **Kipp M.** 2000. *ANVIL. University of Applied Sciences, Augsburg*. www.anvil software.de/index.html [software, disponibile on line].
3. **Sutton V.** 1995. *Lessons in SignWriting: textbook & workbook*. Deaf Action Committee for Sign Writing, La Jolla (CA).
4. **Bianchini C.S.** 2012. *Analyse métalinguistique de l'émergence d'un système d'écriture des Langues des Signes: SignWriting et son application à la Langue des Signes Italienne (LIS)*. Tesi di dottorato, Université Paris VIII & Università degli Studi di Perugia, 512 p.
5. **Cuxac C. & Antinoro Pizzuto E** (2010). *Émergence, norme et variation dans les langues des signes: vers une redéfinition notionnelle*. in: Garcia B., Derycke M. (eds) "Sourds et langues des signes: norme et variations". Langage et société, 131: 37-53.
6. **ELAN Team** (2002). *ELAN - Electronic Linguistic ANnotator*. Nijmegen (Netherlands), Max Planck Institute of Psycholinguistics. www.lat-mpi.eu/tools/elan/ [software, disponibile on line].
7. **Borgia F., Bianchini C.S., Dalle P. & De Marsico M.** (2012). *Resource production of written forms of Sign Languages by a user centered editor, SWift (SignWriting improved fast transcriber)*. Proc. VIII Int. Conf. LREC: 3779-3784.

Uncertain about Uncertainty: Different ways of processing fuzziness in digital humanities data

Binder, Frank
Universität Gießen, Germany

Entrup, Bastian
Universität Gießen, Germany

Schiller, Ines
Universität Gießen, Germany

Lobin, Henning
Universität Gießen, Germany

1 Introduction

The GeoBib project is constructing a georeferenced online bibliography of early Holocaust and camp literature published between 1933 and 1949 (Entrup et al. 2013a). Our immediate objectives include identifying the texts of interest in the first place, composing abstracts for them, researching their history, and annotating relevant places and times. Relations between persons, texts, and places will be visualized using digital maps and GIS software as an integral part of the resulting GeoBib information portal.

The combination of diverse data from varying sources not only enriches our knowledge of these otherwise mostly forgotten texts; it also confronts us with vague, uncertain or even conflicting information. This situation yields challenges for all researchers involved – historians, literary scholars, geographers and computer scientists alike. While the project operates at the intersection of historical and literary studies, the involved computer scientists are in charge of providing a working environment (Entrup et al. 2013b) and processing the collected information in a way that is formalized yet capable of dealing with inevitable vagueness, uncertainty and contradictions. In this paper we focus on the problems and opportunities of encoding and processing fuzzy data.

2 The uncertainty about uncertainty: How to model and represent it

The data collected in our project concerns such different entities as texts, persons and places and is compiled from different sources and different scholarly perspectives. The project is entirely interdisciplinary: besides literary scholars and historians, also geographers and computer scientists are involved. Students and researchers from literary and historical studies are our target audience for whom the resulting online platform shall provide an attractive research tool. Hence, the collected data does not only lie in the intersection of research interests of these fields, but extends to the sum of these interests. The platform is supposed to help answer questions that arise in the field of literature, e.g. finding texts concerned with certain places in a given time period, but also to support historians in finding possible eye witness reports of the crimes of Nazi Germany. Accordingly, information of various kinds is collected with the intention of supporting such diverse use cases. The different scholarly perspectives also determine the amount and kind of data we collect, and their information needs can hardly be covered by a single formalism or predefined ontology. We need a flexible yet coherent formalization that is adaptable to our objectives.

Our workflow and approach to collecting data is one of *divide et impera*: Instead of proposing one format that does it all, we distinguish between different kinds of information depending on the entities concerned. Information collected on the authors of the holocaust texts and relevant places is stored in a user friendly MediaWiki system, while information on these texts is stored in TEI/XML files. Both systems are interconnected and geographical references are integrated as well (cf. Entrup et al. 2013b). The resulting information portal will be backed by an object-relational database.

2.1 Persons and places: Capturing FUZZY information in a Wiki System

Within the field of prosopography the combination of different, possibly contradicting sources is a well-known problem. Pasin and Bradley (2013) offered insights on how such alternative views on historical events could be described using an underlying ontology. Software libraries intended to support processing of prosopographic data are also being developed (e.g. Barabucci and Zingoni, 2013). The GeoBib project collects information and short biographies of authors – a task that bears resemblance to prosopographic research. Many of those authors only published one text. Researching their personal information often leads to ambiguous results, such as different names used, differing information on birth or death dates as well as other personal data.

We extended the MediaWiki system that we use to collect information on persons and places with a set of templates that help to ensure that such information is added in a coherent way, while allowing the data to be vague or apparently contradictory. The Wiki allows the editors to add uncertain information into proposed fields, so that, for instance, different names can be added to one person. Furthermore, the short biographical texts we compose for most of the authors can be used to communicate dissent between different sources.

2.2 TEI/XML: Encoding uncertainty in literary annotations

Literary texts, and as such especially autobiographies and memoires, are not collections of historical facts arranged in an exact chronology. Especially the early Holocaust texts are emotionalizing (cf. Feuchert, 2012, Hickethier 1986) and “conveying the experience made with the National Socialist terror system in a literary, or better, literarised way” (Feuchert, 2012, p. 218). But even apparently factual accounts of events carry a certain degree of vagueness, which leads both historians and literary scholars to interesting research questions, and poses a challenge for data modeling and database design.

Vagueness already occurs when collecting formal metadata. For each holocaust text under consideration we try to identify the first published edition. Yet looking at some more widely known texts of that era, we find multiple editions that differ in such basic information as publisher, year of publication, editor, or even the title. Such phenomena are familiar to those concerned with bibliographic information. Special care is required when formalizing such inconsistent data. In our collection of TEI/XML files, every single edition is represented by one XML file. These files contain the bibliographic information in the TEI header, and they are linked to the corresponding other editions.

TEI provides the @cert attribute for indicating (un)certainty of information given in an XML node. While this strategy allows us to indicate possible vagueness in a machine-readable way, we also need to find ways of communicating this uncertainty to the human user of our information portal. For that purpose, uncertainty regarding the plot or the history of reception of a literary work is captured in literary annotations, i.e. running text, which is an effective and straight-forward way of communicating uncertainty to human readers. In this context, we also allow adding <note> elements to be used by our editors for supplying small texts that will be presented to the user describing the kind of uncertainty involved (cf. Bradley and Short, 2005). The combined use of both elements allows conveying uncertainty to the human user while keeping it encoded in a machine-readable (or machine-traceable) way.

2.3 Modeling Uncertainty in a Database

The GeoBib information portal will rest on a PostgreSQL database (Scherbaum 2009) [4], where we use object relations for the description of certainty and note elements. Our first example represents a person (see Figure 1a). There are three different names associated with the entity: a birth name, the name after her wedding, and a pseudonym.

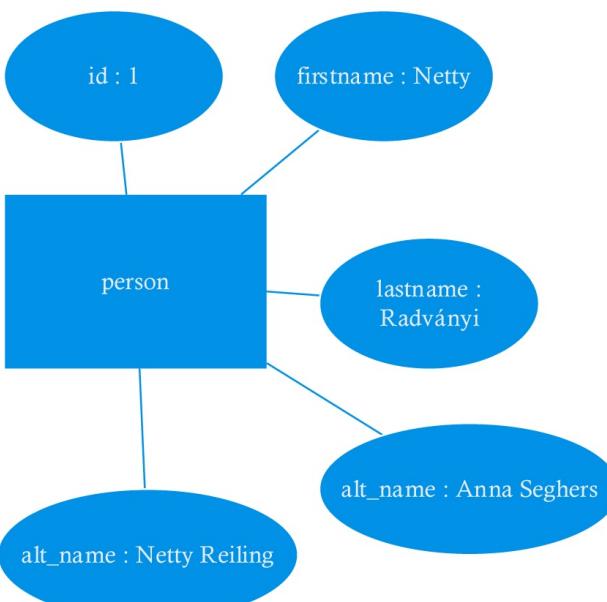


Fig. 1: Exemplary database entries for a person : a) person entity model

person	
PK	id : 1
	firstname : Netty
	lastname : Radványi

Fig. 2: Exemplary database entries for a person : b) simplified perso object

oid_alt		oid_alt	
PK	OID : 1	PK	OID : 1
PK	tablename : person	PK	tablename : person
PK	column : name	PK	column : name
PK	alt : Netty Reiling	PK	alt : Anna Seghers
	certainty : high		certainty : high
	note : NULL		note : NULL

Fig. 3: Exemplary database entries for a person : c) related alternative/additional information

While a person entity has certain fields that can be filled in (see Fig. 1b), we use object relations (Fig. 1c) to add alternative information and related values of certainty and/or a note.

The second example describes a literary work with an uncertain year of publication. A relational database would require intermediate tables for all attributes of one entity, which may have uncertainties and/or notes attached (cf. Bradley and Short, 2005). In an object relational database using one special entity is sufficient in such a case.

werk		oid_alt	
PK	id : 1	PK	OID : 1
	title : KZ Sachsenhausen	PK	tablename : werk
	publisher : Lucie Großer	PK	column : pub_date
	pub_date : 1949	PK	alt : NULL
	extent : 39		certainty : low
			note : NULL

Fig. 4: Work entity and related certainty field "pub_date"

As shown in Fig. 2, the table includes the object ID, the table name, the relevant column name, and the alternative content. Each dataset may contain a certainty attribute and/or a note. The certainty field is defined, in accordance with TEI, as either {high, medium, low, unknown} and the note field is a text that will be presented to the user and is meant to explain the uncertainty[6]. In the example above (Fig. 2) a year of publication is given but its certainty is marked as "low". Such relations can be added for every field of every entity in the database.

3 Discussion and Outlook: Surely more uncertainty

We have just presented our approach of encoding uncertainty in our database. Such information can be used, for instance, to rank search results or to increase recall on certain queries and parameter settings. But we still see more challenges ahead: The encoded uncertainty has to be communicated effectively to the human user. Accordingly, the visualization of uncertainty, and especially the presentation of search results based on uncertain or divergent information will be among our next concerns. A similar problem of visualizing uncertainty arises in the forthcoming georeferencing and geotagging: Literary texts are no geographical maps. They constitute themselves in their geographical space but might encode this information in a way hard to decipher (cf. Reuschel et al., 2013). Fictional place names can sometimes be identified with actual places on a map, but sometimes it is impossible to do so. Geographical locations may be referred to by metaphors, old or forgotten names, local identifiers or nicknames. Such informal references frequently remain geographically imprecise and require interpretation (cf. Hill, 2006, p. 28f). The textual material that our editors provide could also be used to test and improve automatic methods of geographical relation extraction (e.g. Blessing and Schütze, 2010). Still, automatic georesolving is hindered by the limited (historical) coverage of contemporary gazetteers, spelling changes and changing administrative boundaries (Tobin et al., 2010, p. 8). Such limitations also pertain to prisons and camps, those places of high interest in our domain, whose exact geographical locations have to be reconstructed manually before adding them to our databases.

4 Acknowledgements

Funded by the German Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung, BMBF) from July 2012 to June 2015 (FKZ: 01UG1238-A-B).

References

- Barabucci, Gioele and Jacopo Zingoni** (2013). PROSO: prosopographic records. In: *Proceedings of the 1st Workshop on Collaborative Annotations in Shared Environments: metadata, vocabularies and techniques in the Digital Humanities*, DH CASE '13. September 10 2013, Florence, Italy <doi:10.1145/2517978.2517982>

Blessing, Andre and Hinrich Schütze (2010). *Self-annotation for fine-grained geospatial relation extraction*. In: Proceedings of the 23rd International Conference on Computational Linguistics pp. 80-88. dl.acm.org/citation.cfm?id=1873781.1873791

Bradley, John, and Harold Short (2005). *Texts into Databases: The Evolving Field of New-Style Prosopography*. In: Literary and Linguistic Computing, 20 (2005), 3-24 <doi:10.1093/llc/fqj022>

Entrup, Bastian, Maja Bärenfänger, Frank Binder and Henning Lobin (2013a): *Introducing GeoBib: An Annotated and Geo-referenced Online Bibliography of Early German and Polish Holocaust and Camp Literature (1933–1949)*. Digital Humanities 2013, University of Nebraska-Lincoln, 16-19 July 2013. dh2013.unl.edu/abstracts/ab-229.html

Entrup, Bastian, Frank Binder and Henning Lobin (2013b): *Extending the possibilities for collaborative work with TEI/XML through the usage of a wiki-system*. In: Proceedings of the 1st Workshop on Collaborative Annotations in Shared Environments: metadata, vocabularies and techniques in the Digital Humanities, DH CASE '13. September 10 2013, Florence, Italy. <doi:10.1145/2517978.2517988>

Feuchert, Sascha (2012). *Fundstücke: Bemerkungen zu Darstellungskonventionen und paratextuellen Präsentationsformen früher Texte deutschsprachiger Holocaustliteratur*. In: Günter Butzer / Joachim Jacob (Hg.): Berührungen. Komparatistische Perspektiven auf die frühe deutsche Nachkriegsliteratur. München: Wilhelm Fink 2012, pp. 217-230.

Hickethier, Knut (2006). *Biographie, Autobiographie, Memoirenliteratur*. In Ludwig Fischer (eds), Literatur in der Bundesrepublik bis 1967. München 1986, pp. 574-584.

Hill, Linda L. (2006). *Georeferencing*. The Geographic Associations of Information, Cambridge: The MIT Press.

Pasin, Michele, John Bradley (2013). *Factoid-based Prosopography and Computer Ontologies: towards an integrated approach*. In: Literary and Linguistic Computing (2013). <doi:10.1093/linc/fqt037>

Reuschel, Anne-Kathrin, Barbara Piatti and Lorenz Hurni (2013). *Modelling Uncertain Geodata for the Literary Atlas of Europe*. In: K. Kritz et al. (eds.) Understanding Different Geographies. Lecture Notes in Geoinformation and Cartography, <doi:10.1007/978-3-642-29770-0_11>

Scherbaum, Andreas (2009). *PostgreSQL – Datenbankpraxis für Anwender, Administratoren und Entwickler*. Open Source Press. München, 2009.

Tobin, Richard, Claire Grover, Kate Byrne, James Reid and Jo Walsh (2010). *Evaluation of georeferencing*. In: Proceedings of the 6th Workshop on Geographic Information Retrieval, GIR '10, Zurich, Switzerland, <doi:10.1145/1722080.1722089>

As of December 2013 the GeoBib project has identified and collected bibliographical information on 670 texts of early Holocaust and camp literature and produced 130 annotation documents so far. These include references to 620 authors, 550 locations, 230 camps and 45 ghettos.

See for instance svario.it/factoid

German original: „*Bereits seit 1933 sind, vor allem im Ausland, Texte erschienen, die Erfahrungen mit dem nationalsozialistischen Terrorsystem literarisch oder besser: literarisert vermitteln*“ (Feuchert, 2012, p. 218)

<http://www.postgresql.org/about/>

We do not use the PostgreSQL entity parameter “WITH OIDS”, because of the high memory requirements (cf. Scherbaum 2009, p. 161f). Our OID (object identifiers) are composed of the entity id and name.

Alternatively, the note field could also contain additional information.

katherine.bode@anu.edu.au
Australian National University

1. Introduction

From colonial times to World War Two, most of Australia's many newspapers incorporated serial fiction, including local and overseas titles. The Australian fiction in these periodicals has largely been identified (Austlit), and important research in this area is ongoing (Bode 2012; Gelder 2011). However, very little is known about the overseas works, including the titles, authors and themes, and their circulation and reception in Australia.

An important reason for this lack of knowledge is the size of the archive. With hundreds of newspapers – many containing multiple instalments of novels per edition – a systematic manual search for fiction is unfeasible. The search possibilities for this archive have dramatically expanded with the creation of the National Library of Australia's (NLA) Trove database. From 2007 to 2012, the NLA digitised over four million pages of Australian newspapers, from every state and territory, published from 1803 to the 1950s. Combined with digital humanities methods for data mining and analysis, this ongoing digitisation project makes identifying serial fiction in Australian newspapers possible for the first time in a systematic and reliable way.

The project reported in this paper describes a computer-enabled approach to exploring the presence, circulation and reception of fiction in Australian newspapers that enables research, and advances arguments, relevant to bibliographical, book historical and literary studies as well as digital humanities.

2. Bibliography

The project showcases how digital humanities methods can significantly enhance bibliographical records and knowledge. Searching Trove using terms associated with serial fiction – including ‘chapter’, ‘story’ and ‘fiction’ – enables identification of potentially relevant records. The bibliographic information and full text results of these searches are extracted as CSV and text files using a Python harvesting tool developed by Sherratt (2013). These files are supplemented through additional research (for instance, the authors' nationalities and gender), and transfer to a database that will be freely accessible to researchers and the public.

This approach is providing extremely effective in identifying serial fiction. The first search – of ‘chapter’ – yielded approximately 200,000 individual instances of fiction in Australian newspapers. Many other searches remain to be done; however, even this initial result demonstrates this method's capacity to enhance bibliographical records. This search process will undoubtedly reveal previously unrecorded instances of publication, particularly of non-Australian fiction. Some of these instances will almost certainly be of titles that have not been indexed previously, including by well-known authors. More broadly, this project demonstrates the potential of digital humanities methods to maximise the utility, and thus enhance the value and consequence, of digital collections.

3. Book History

The collected bibliographic data enables quantitative analysis of the transnational movement of fiction. This approach builds on earlier studies, most prominently, Moretti's ‘distant reading’ (2005) and, more recently, Jockers's ‘macroanalysis’ (2013). In terms of the archive searched and the cultural phenomena analysed, it is also indebted to Nicholson's identification and analysis of American jokes in digitised nineteenth-century British newspapers (2012). Importantly, however, unlike these other works, the body of data underpinning this project's arguments and findings will be publicly available, so other scholars can explore, check, extend and potentially challenge the findings; and so this data can be reused in future research.

Mining a 'Trove': Modelling a Transnational Literary Culture

Bode, Katherine

Findings of initial data analysis, for 1830 to 1880, already indicate trends that challenge existing perceptions of Australian literary culture. Where metropolitan newspapers are routinely identified as the main Australian serial fiction publishers (e.g. Webby 2000), this study highlights the strong involvement of regional newspapers. This finding challenges the existing centre/periphery understanding of colonial literary culture. Also contesting this model is the revelation that – while overseas fiction has been estimated to vastly outnumber local titles (Morrison 1998) – in this period, more local than overseas fiction was published. One interesting outcome of this strong local publication is a reversal of the much-discussed female-dominance of nineteenth-century serial fiction authorship. Although most American and British serial fiction was by women (Casey 1996; Coultrap-McQuin 1990; Thompson 1999), men wrote the majority of titles in Australian periodicals in this period. While local titles outnumbered overseas fiction, this initial search has identified a significant amount of non-Australian titles, including a higher-than-anticipated number of American stories, as well as fiction from a wide range of countries besides Britain, including China, Russia, France and Germany. As well as highlighting the status of Australian periodicals as ‘contact zones’ (Pratt 1990) for literature, this range of national literatures further challenges a centre/periphery understanding of colonial literary culture.

This project’s combination of digital humanities and book history suggests important directions for the former as well as the latter field of study. Book history is increasingly recognised as playing an important role in the development of digital humanities. Alan Liu describes book history as a Levi-Straussian ‘trickster figure’ for digital humanities, uniting the field’s commitment to older humanities disciplines, and the value of the old itself, with more recent interest in emergent media and design (2013, 410). Elsewhere he points to the way book historians ‘increasingly compare, and not just contrast, earlier writing/reading practices to their digital successors’ (2012, 16), and the potential of this approach to enhance understanding the digital age and the digital humanities.

The project employs this comparative framework to consider reading practices. While one might assume nineteenth-century newspapers differ entirely from the Internet, in fact both are networked interfaces uniting various content, including that previously published elsewhere, for readers who have significant autonomy in deciding what to read and what connections to draw. Notwithstanding these significant parallels, it is equally important that the use of digitised archives, and digital humanities search and retrieval methods, not occlude historical context. In particular, this project works to maintain a view of nineteenth-century newspapers as coherent and interconnected cultural artefacts rather than containers of discrete content (a perception potentially encouraged by search results in the form of individual articles).

4. Literary Studies

The full-text records extracted from Trove provide the basis for computer-assisted textual analysis, particularly topic modelling. This aspect of the project will follow, and in so doing, test and extend Jockers’s analysis of influence in relation to Irish, English and American literature (2013). Topic modelling will be used to investigate whether, and if so, to what extent, local stories in Australian newspapers employed similar themes, language, or generic strategies to the other-national literatures alongside which they were published. The same method will be used to consider relationships between other-national literary forms. Like Moretti’s and Jockers’s analyses, this project will contribute to shifting literary studies beyond a nation-based framework. However, where these earlier studies consider general bibliographic corpora, in exploring texts published alongside one another, this project provides an important opportunity to consider influence in relation to a specific material context: that is, fiction received and experienced by particular readers at particular times.

5. Digital Humanities

McCarty’s notion of modelling is a key concept in this project’s formulation and development. In McCarty’s words, a model is ‘an abstraction or simple representation of a more complex real phenomena’ (2008), and modelling enables exploration of and experimentation with phenomena that would otherwise be intractable or inaccessible (2005: 27). This project will complicate and extend this methodological framework by highlighting the multiple number and layers of models and modelling processes involved in exploring serial fiction in Australian periodicals. These layers include the digitised newspaper pages (themselves created from other models, predominantly microfiche), the Trove database more broadly, the database in which the search and harvesting results are represented, as well as the subsequent quantitative analyses of bibliometric and textual data. Where McCarty has always insisted upon the status of models as fictions, this foregrounding of multiple and layered models emphasises the radical contingency of this foundational concept for digital humanities, as well as the theoretical nature of its outcomes.

Foregrounding the contingent and theoretical nature of modelling has two key implications for this project, and for digital humanities research broadly. First, it provides the groundwork for working with an historical record that necessarily contains multiple gaps: Trove has not digitised all Australian newspapers; some records have been lost, others are still to emerge; the quality of OCR for the texts differs radically; and the search process will not discover all serial fiction in Trove. Second, it enables a recognition that even the historical record we have – including what might be considered its obvious facts – needs to be treated as contingent and theoretical. For instance, bibliographic details added to the database – such as the name and gender of authors – are obviously facts, but may not have been present to historical readers (stories were published anonymously or under pseudonyms) and thus cannot be taken as absolute points of reference for understanding the historical circulation and reception of fiction. In moving away from understanding quantitative analysis of archival records as proof of historical phenomena, the underlying framework seeks to forge a conversation between bibliographers, archivists, book historians, literary critics and digital humanists that is data-rich, but oriented towards theoretical possibilities and constructs rather than proof and measures.

References

- Austlit: The Australian Literature Resource.* (2002–). www.austlit.edu.au.
- Bode, K.** (2012). *Reading by Numbers: Recalibrating the Literary Field*. London: Anthem Press.
- Casey, E.** (1996). Edging Women Out? Reviews of Women Novelists in the *Athenaeum*, 1860–1900. *Victorian Studies* 39.2: 151–71.
- Coultrap-McQuin, S.** (1990). *Doing Literary Business: American Women Writers in the Nineteenth Century*. Chapel Hill: University of North Carolina Press.
- Gelder, K.** (2011). Negotiating the Colonial Australian Popular Fiction Archive. *JASAL* Special Issue: Archive Madness: 1–12.
- Jockers, M.** (2013). *Macroanalysis: Digital Methods and Literary History*. Champaign: University of Illinois Press.
- Liu, A.** (2012). The State of the Digital Humanities: A Report and a Critique. *Arts and Humanities in Higher Education* 11.1–2: 8–41.
- Liu, A.** (2013). The Meaning of the Digital Humanities. *PMLA* 128.2: 409–23.
- McCarty, W.** (2005). *Humanities Computing*. London: Palgrave Macmillan.
- McCarty, W.** (2008). Knowing ...: Modeling in Literary Studies. In Susan Schreibman and Ray Siemens (eds), *Companion to Digital Literary Studies*. Oxford: Blackwell. <http://www.digitalhumanities.org/companionDLS/>.
- Moretti, F.** (2005). *Maps, Graphs, and Trees: Abstract Models for Literary History*. London: Verso.

- Morrison, E.** (1998). Serial Fiction in Australian Colonial Newspapers. In John O. Jordan and Robert L. Patten (eds), *Literature in the Marketplace: Nineteenth-Century British Publishing and Reading Practices* (2nd ed.). Cambridge: Cambridge University Press, pp. 306-24.
- National Library of Australia.** (2007-). Trove Database. <http://trove.nla.gov.au/>.
- Nicholson, B.** (2012). 'You kick the bucket; we'll do the rest': Jokes and the Culture of Reprinting in the Transatlantic Press. *Journal of Victorian Culture* 17.3: 273-86.
- Pratt, M. L.** (1991). Arts of the Contact Zone. *Profession*: 33-40.
- Sherratt, T.** (2013). Mining the Treasures of Trove (part 1). Discontents. Blog. <http://discontents.com.au/mining-the-treasures-of-trove-part-1/>.
- Thompson, N.** (1999). Responding to the Woman Questions: Rereading Noncanonical Victorian Women Novelists. In Nicola Diane Thompson (ed.), *Victorian Women Writers and the Woman Question*. Cambridge: Cambridge University Press, pp. 1-23.
- Webby, E.** (2000). Colonial Writers and Readers. In Elizabeth Webby (ed.), *The Cambridge Companion to Australian Literature*. Cambridge: Cambridge University Press, pp. 50-73.

Distant reading of naïve poetry: corpora comparison as research methodology

Bonch-Osmolovskaya, Anastasia

National Research University Higher School of Economics Moscow, Russian Federation

Orekhov, Boris

National Research University Higher School of Economics Moscow, Russian Federation

The method of distant reading - term proposed by Franco Moretti (Moretti 2005 , Moretti 2013) , or – using another term – macroanalysis (Jockers 2013) has become a mainstream in digital humanities in last couple of years. The general idea of the approach is to gain new knowledge about literary and cultural processes with the help of digital and quantitative models applied to all sorts of language or literary resources. In our paper we follow the distant reading method focusing on the phenomenon of naïve poetry – poetical opuses, composed by non-professional poets and distributed on special web- sites. One of them – Russian stihi.ru (*stihi* means 'verses') – has nowadays become a giant collection of dilettante literature with more than 5000 authors, about 21 mln of works and everyday update. It, thus, can be regarded as an extraordinary cultural linguistic resource made by crowd sourcing. At the same time Russian National Corpus resources possess a unique resource - Poetic corpus (Grishina et al. 2009; <http://ruscorpora.ru/search-poetic.html>). In contrast to stihi.ru, Russian Poetic Corpus has been collected and marked up by the team of experts in linguistic and literary studies and it presents Russian poetical classics from the 18th century till the early 1930th.

Comparison of the two poetic resources of naïve and classical poetry gives us an excellent possibility to use quantitative analysis to get some promising insights. We can understand more about the nature of literary imitations and epigone writings, the foundations and circulations of literary canon, the mechanisms of prosaic/poetic language shifts and many other topics related to sociology of textual culture that could not be studied with traditional methods. Some steps taken in this direction are presented in this paper.

We look first of all at frequency measures in both resources and analyze the revealing fluctuations of different word frequencies. We use the frequency list of Russian National Corpus (called below the general frequency list) as a controlling benchmark, that helps us to separate the words, which are most frequent in common Russian Lexicon, from those, which get high frequency exclusively in Russian verses, high or naïve.

Then we make a qualitative analysis of the nouns that occur in the top 100 of each frequency list. We identified a range of semantic domains that can be expressed by these nouns and compared the domains of each poetical resource. As a result we defined three main strategies of naïve composition.

Preparation

For our research we have taken a sample of 50 mln word usage from naïve poetic corpus, which makes a representative corpus of more than 54 thousand authors. As our main aim was to find out what poetic patterns from high poetry are apt to be borrowed and imitated, we decided to extract a subcorpus of most typical imitating poetry. We searched for the authors who would bear in mind high poetical examples and try to go with them in their composition. The sorting has been made automatically. First we extracted all the bigramms from the high poetical corpus, and then we took only those documents from the naïve corpus, which a) have at least 50% bigrams that coincide with the bigrams of the high poetry list, b) are longer than 20 bigramms. Our final sample consisted of almost 9 mln word usage. The high poetic corpus has about 8 mln word usage. We lemmatized all the words in both corpora. After lemmatization the naïve poetic corpus consisted of 84 thousand lemmas

Methodology

We conducted three analyses based on the comparison of the three frequency lists: the frequency list from naïve poetic sample, the frequency list from Russian Poetic Corpus and the general Russian frequency list based on Russian National Corpus. In the first experiment we compared the change of ranks of very high frequent words in the naïve sample relatively to high poetic and general lists. Secondly we considered outliers of naïve poetry frequency list: those words that demonstrate dramatically different frequency behavior. The last experiment consisted in relating all the top 100 nouns of each frequency list to semantic domains, that they are most probably used for. Then the contents and the variety of each domain in each list has been analysed.

Results

The interpretation of the resource comparison results can be summarized by defining three basic strategies in naïve poetry: imitation, self-actualization and naming. Each of the strategies will be illustrated below by data examples.

1. The top frequency list of naïve poetical resource shows interesting deviation both from high poetical corpus list and general frequency list (see table 1)

Word (in Russian)	Word (translation)	Position rank in naïve list	Position rank in high poetical list	Position rank in general frequency list
и	and	1	1	1
я	I	2	3	5
не	not	3	4	3
в	in	4	2	2
ты	you	5	7	33
то	this	6	5	23
что	that	7	11	9
быть	be	8	10	6
на	on	9	6	4
как	as	10	9	19
с	with	11	8	8
мы	we	12	13	18
а	but	13	17	10
мой	my	14	15	60
но	but	15	14	16
так	so	16	27	30
за	for behind	17	22	24
любовь	love (noun)	18	52	307
любить	love (verb)	19	66	181

As we can see from the table, the naïve poetry demonstrates important lexical features, some of them are specific, and some of them are typically poetic, being shared with the list of high poetry frequency. We observe an interesting tendency at the very top, where personal pronouns I and you displace the most common propositions in and on from the second and the forth positions correspondingly. Both pronouns I and you can be considered as lexical traits of poetical discourse. But the naïve poetry shows higher ranks for both of them (2 vs 3 in high poetry, and 5 vs. 7 in high poetry correspondingly). We see increase of frequency of those words which are already indicative for high poetical frequency list. The tendency to intensify specific poetical lexical features can be called the imitative strategy. The rank shift of the words love (noun and verb) is even more straightforward manifestation of the same strategy. While in general frequency list those words are not even in top 100, they occupy 52 and 66 positions in high poetical list and so far being an etalon of poetical shift they become the most frequent words in naïve poetry.

2. The table below shows 5 nouns which have the biggest difference of ranks between naïve poetry list noun frequency of the Russian Poetical corpus (see table 2)

word	word (translation)	Position rank in naïve list	Position rank in high poetical list	Position rank in general frequency list
фото	photo	971	30827	3400
сигарета	cigarette	939	28957	2109
проблем	problem	610	12572	197
мама	mum	330	2125	309
девчонка	babe, gal (derivative from girl)	865	5030	2563

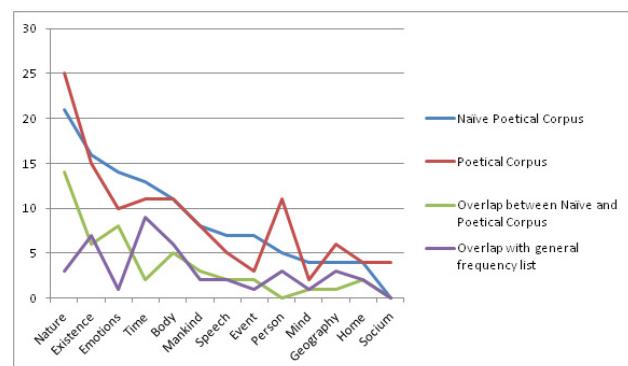
These words behavior is various: some of them are more frequent comparing the general frequency list (*photo, cigarette*), some of them stay on roughly the same rank (*mum*), some are more rare than in general, but still show immense difference

with the high poetical list (*problem*). The gap between naïve and poetic frequency list signals that there are some semantic zones where naïve poetry seems to be independent from the classical poetic canon. This trend can be defined as a self-actualization strategy which is in some sense opposite to the imitative strategy.

3. We took top 100 nouns of every list and compared their lexical distribution. The nouns had been grouped into abstract semantic domains. Some words could be associated with several domains due to their polysemy. As a result we have identified 13 semantic domains, 12 of them are shared between naïve poetry, high poetry and common frequency lists and the 13th is not presented in the naïve poetry list. The domains we have defined are as follows:

Mankind (everything that may characterize a person: *soul, beauty, name, heart, strength* etc.), Body, Emotions, Mind, Existence (*God, world, truth, time, fate* etc.), Speech, Person (*father, son, friend, enemy* etc.), Event (*love, happiness, past, disaster* etc.), Time, Nature, Geography (*road, hill* etc.), Home (*window, door* etc.). The 13th domain which is found only in the high poetical list and in the general frequency list is Social and it includes such words as people, labor, fame in the poetical list, and state, money, law etc. in the general list.

Analysis of the overlaps and varieties of the naïve and high poetical lists showed differences in the elaboration of the domains in two corpora. The general frequency list helped to draw out the words that are commonly frequent and their presence in the list cannot be understood as a signal of the poetic concentration on the domain. The results are demonstrated on the graph below:



As we can see from the graph, there are three zones of the naïve poetical sample that demonstrate high lexical variety of frequent nouns in comparison to the poetical corpus. These are Emotions, Event and Speech. Most of the words of those domains are not frequent in general lexicon. The lexical multiplicity can be explained by extensive strategy: the naïve poets do not use sophisticated verbal apparatus to express the conceptual space of the verse, but prefer straightforward lexical naming (*pain, wish, encounter, grief, love, question, answer* etc.)

References

Grishina E., Korchagin K., Plungian V., Sitchinava D. *Poeticheskiy korpus v ramkah NKRYa: obshchaya struktura i perspektivy ispol'zovaniya 'Natsional'nyy korpus russkogo jazyka: 2006-2008. Novye rezul'taty i perspektivy'*. Saint Petersburg, 2009. P. 71-113. [Poetic Corpus in RNC: general structure and using perspectives]

Moretti, Franco. *Graphs, Maps, Trees: Abstract models for a literary history*. Verso, 2005.

Moretti, Franco. *Distant Reading*. Verso, 2013

Jockers, Matthew L. *Macroanalysis*. University of Illinois Press, 2013

Dimensions of literary appreciation. Word use and ratings on a book discussion site

Boot, Peter

peter.boot@huygens.knaw.nl
Huygens ING, Netherlands, The

Introduction

The appreciation of literature is a subjective process. In reading and judging books, characteristics of individual readers interact with characteristics of books and their reputation. This paper looks at book ratings on a book discussion site and tries to assess the role of individual readers' characteristics in these ratings. For that purpose, the paper inspects on the one hand the textual properties of the review texts that readers contribute to this site, and on the other hand the ratings that they assign.

Given the well-established connection between word use frequencies and authorial style (e.g. Burrows, 2002; Burrows, 2003), the paper hypothesizes that these same style markers in texts by readers will correlate with these readers' quality judgments about books. Patterns in word usage are known to reflect aspects of readers' psychological make-up (Argamon et al., 2005; Noecker et al., 2013; Pennebaker et al., 2003), and these psychological properties, e.g. the Big Five personality dimensions, are related to aesthetic preferences in many fields (Golbeck and Norris, 2013; Gridley, 2013; Zweigenhaft, 2008), including books and literature (Cantador et al., 2013; Wiersema et al., 2012).

Aesthetic appreciation has been shown to be a multi-faceted process (Myszkowski et al., 2014; Rentfrow et al., 2011). Here, I assume that literary appreciation is influenced by multiple aspects of the reader's psychology, such as, among others, his/her cognitive, affective and social dispositions. Therefore, besides investigating the over-all most frequently used words, as stylometry often does, I will also look at the high frequency words within the categories of cognitive, affective, and social words, as defined by LIWC (Pennebaker et al., 2007). I expect that the relative frequencies of e.g. individual social words (rather than the category frequencies that LIWC-based research typically uses) will capture to some extent the nature of a person's sociability and will to that extent also reflect how that sociability affects literary preferences.

Data

The data for this paper come from Dutch book discussion site watleesjij.nu (whatareyoureading.now). The site is similar to e.g. Goodreads, LibraryThing or lovelybooks.de: users rate, label and review books, they can evaluate reviews by others, can strike up friendships with and send messages to other users. I downloaded the site's content in June 2013. I investigate review texts and ratings contributed by the top 20 (in terms of total review length) contributors to the site (I removed two accounts that seemed to be used by multiple persons.) For each of these users, I create a file containing all of the review texts this user has contributed to the site. The average word count is 44036. I also collect the ratings (in terms of one to five stars) for all of the 624 books that were rated by at least two of the twenty users.

Method and results

As a first step, I compute correlations between the word use frequencies in each of the word categories and the book ratings. The word frequencies are represented as a matrix of zscores, where users are rows and words are columns. For the computation of the zscores I use Eder and Rybicki's style script (2011), then select only those words that form part of the relevant LIWC category. The ratings are given in a matrix with users as rows and books as columns. Non-rated books

are represented by 0. To assess the correlation between these matrices I rely on the (bias corrected) distance correlation and the associated significance test as described by Székely and Rizzo (2013). Table 1 reports the results, including the number of words that gave the best results for each category (However, for all categories except Affect the correlations were significant at the .01-level even for the top 25 words.) The table also gives the percentage of words belonging to the category in the review files.

Table 1. Bias-corrected distance correlation between word usage and book appreciation for different word categories

Category	bias corrected distance correlation	p-value	Optimum number of most frequent words	percentage of words in category
All words	.49	<.0001	2900	100
Function words	.36	<.0001	250	50
Affect words	.21	.0025	225	3
Cognitive words	.35	<.0001	375	18
Social words	.47	<.0001	125	12

The table shows that frequencies in each of the word categories are quite significantly correlated with the word ratings. The relatively low effect from affective words may be due to the low percentage of affect words in the texts.

The most striking result is no doubt the performance of the category of social words. In order to further investigate this effect, users were clustered in two groups, based on their usage of social words (I employed the pam partitioning function in R.) I then looked at contrastive word use of these clusters and at the books liked by the cluster members. The oppose function from Eder and Rybicki (2011) was used to find words preferred by either cluster. The results are given in table 2. The first cluster shows an interest in people and especially family that the second cluster, with its mostly cognitive or procedural interest, seems to lack entirely.

Table 2. Words preferred (from all words) by the two clusters (translated from Dutch). For cluster one, only the top 20 preferred words are given

Cluster	Preferred words
1	daughter, parents, family [nuclear], mother, woman, together, father, children, past, young, child, debut, house, brother, women, tells, love, marriage, family [extended], care
2	so, perhaps, page, of course, pity, well, read, precisely, actually, just, immediately, think, for instance, part, viz., believe, even, sort of, interesting, by the way

In order to find out the sort of books preferred by the clusters, I summed the book ratings by cluster. I then selected and diagrammed a subset of books, consisting of the ten books best liked by either cluster, the ten books best liked 'contrastively' by either cluster (computed by subtracting the ratings for cluster 1 from those for cluster 2), and the ten books best liked by both. After removal of duplicates, thirty books remained. Figure 1 displays the books with their ratings by the two clusters. Point and title size reflect popularity on the site. Grayscale represents

genre. Point positions were slightly changed to avoid overlap. Lines between title labels and points were suppressed in the interest of clarity.

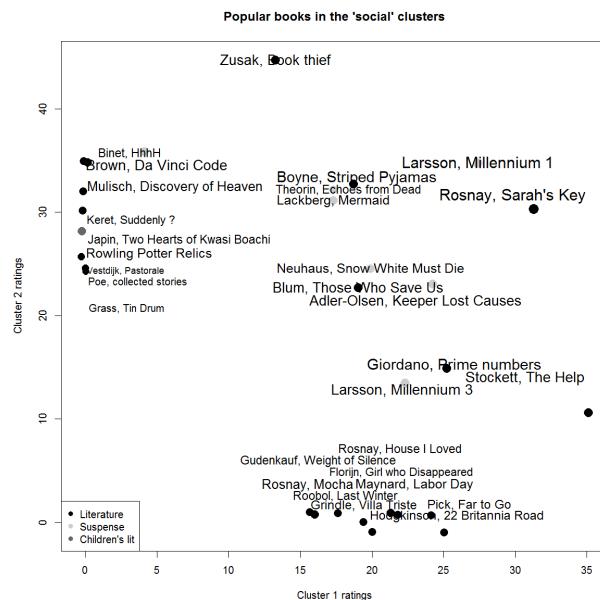


Fig. 1: Books as rated by the two clusters.

The figure seems to show some systematic differences in the preferences of the two clusters. Cluster 1, that uses mostly family-oriented words, seems to prefer slightly more popular books (larger point size). Cluster 2, that uses procedural or cognitive words, has strong preferences for a number of staunchly 'literary' works, such as those by Grass, Binet and classical Dutch authors. As to a potential preference for suspense novels, this figure does not allow us to draw any firm conclusions.

Discussion

The reason why different people prefer different books has often been sought in differing literary norms (e.g. Von Heydebrand and Winko, 2008). This explanation is not quite satisfactory, for two reasons: first because it does not explain why people develop different norms, and second because there are no *a priori* reasons why norms rather than, say pleasure or 'thrills' (Konecni, 2005) should determine one's preference for one book over another. This paper takes another approach and the results presented here tentatively establish the existence of a correlation between book preferences and patterns of word usage in several psychologically meaningful categories. Especially the relation between the pattern of usage of social words and literary appreciation seems very strong, confirming the importance of extraversion for aesthetic judgment noted by Furnham and Chamorro-Premuzic (2004), but appreciation is also clearly related to usage of cognitive words and of function words.

There are some obvious limitations to this experiment. The number of subjects is very small (dictated by the need to have a sufficient number of words). It would also have been better to use texts from another domain. However, an exploratory analysis of the effect of clustering based on social word usage seems to show that the verbally more 'social' group prefers the less literary or more popular novel. Given the small numbers, more than provisional results should perhaps not be expected.

Next steps should include clustering on the basis of other word categories, an investigation into the independent effect of these categories, and case studies at the level of individual readers. It would also be very interesting to see to what extent the literary norms that readers formulate in the reviews can be shown to be related to the word usage patterns as discussed here.

References

- Argamon, S., S. Dhawle, M. Koppel and J.W. Pennebaker.** (2005) 'Lexical predictors of personality type', Proceedings of the Joint Annual Meeting of the Interface and the Classification Society of North America.
- Burrows, J.** (2002) "Delta": A measure of stylistic difference and a guide to likely authorship. Literary and Linguistic Computing 17(3): 267-287.
- Burrows, J.** (2003) 'Questions of Authorship: Attribution and Beyond. A Lecture Delivered on the Occasion of the Roberto Busa Award ACH-ALLC 2001, New York'. Computers and the Humanities 37(1): 5-32.
- Cantador, I., I. Fernández-Tobías, A. Bellogín, M. Kosinski and D. Stillwell.** (2013) 'Relating Personality Types with User Preferences in Multiple Entertainment Domains', Proceedings of the 1st Workshop on Emotions and Personality in Personalized Services (EMPIRE 2013), at the 21st Conference on User Modeling, Adaptation and Personalization (UMAP 2013).
- Eder, M. and J. Rybicki.** (2011) 'Stylometry with R'. Paper presented at Digital Humanities 2011: Conference Abstracts, Stanford University, Stanford, CA.
- Furnham, A. and T. Chamorro-Premuzic.** (2004) 'Personality, intelligence, and art'. Personality and Individual Differences 36(3): 705-715.
- Golbeck, J. and E. Norris.** (2013) 'Personality, movie preferences, and recommendations', Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining: ACM, pp. 1414-1415.
- Gridley, M.C.** (2013) 'Preference for Abstract Art According to Thinking Styles and Personality'. North American Journal of Psychology 15(3).
- Konecni, V.J.** (2005) 'The aesthetic trinity: Awe, being moved, thrills'. Bulletin of Psychology and the Arts 5(2): 27-44.
- Myszkowski, N., M. Storme, F. Zenasni and T. Lubart.** (2014) 'Is visual aesthetic sensitivity independent from intelligence, personality and creativity?'. Personality and Individual Differences 59(16-20).
- Noecker, J., M. Ryan and P. Juola.** (2013) 'Psychological profiling through textual analysis'. Literary and Linguistic Computing 28(3): 382-387.
- Pennebaker, J.W., R.J. Booth and M.E. Francis.** (2007) 'Linguistic Inquiry and Word Count (LIWC2007)', Linguistic Inquiry and Word Count (LIWC2007). Austin, TX.
- Pennebaker, J.W., M.R. Mehl and K.G. Niederhoffer.** (2003) 'Psychological aspects of natural language use: Our words, our selves'. Annual review of psychology 54(1): 547-577.
- Rentfrow, P.J., L.R. Goldberg and R. Zilca.** (2011) 'Listening, watching, and reading: The structure and correlates of entertainment preferences'. Journal of personality 79(2): 223-258.
- Székely, G.J. and M.L. Rizzo.** (2013) 'The distance correlation t-test of independence in high dimension'. Journal of Multivariate Analysis 117(193-213).
- Von Heydebrand, R. and S. Winko.** (2008) 'The qualities of literatures', The Quality of Literature: Linguistic Studies in Literary Evaluation. Amsterdam: Benjamins, pp. 223-239.
- Wiersema, D.V., J. Van Der Schalk and G.A. van Kleef.** (2012) 'Who's afraid of red, yellow, and blue? Need for cognitive closure predicts aesthetic preferences'. Psychology of Aesthetics, Creativity, and the Arts 6(2): 168.
- Zweigenhaft, R.L.** (2008) 'A do re mi encore: A closer look at the personality correlates of music preferences'. Journal of individual differences 29(1): 45.

An XML Schema to Interpret Networked Biographies: Reading Mid-Range

Booth, Alison

University of Virginia, United States of America

Martin, Worthy

University of Virginia, United States of America

Collective Biographies of Women, is an open-access project supported by the Institute for Advanced Technology in the Humanities, Scholars' Lab, and the English Department at the University of Virginia, as well as an ACLS Digital Innovation Fellowship. In recent years it has grown from an online bibliography of all English-language books that collect three or more short biographies of women into a digital prosopography that interrelates women, printed books, and narratives in what we call documentary social networks (introduced at DH 2013). CBW stands out as a literary study of prosopographies in the print era, and primarily the transatlantic nineteenth century (see the bibliography, <http://womensbios.lib.virginia.edu>). Most research that employs the term *prosopography* allies itself with history or classical and medieval studies, and today, relies on databases and websites. We work with the concept as it is often defined, as collective biography, that is, printed prose collections of short biographies (see the selective bibliography for a context on prosopography, nonfiction narrative, and our method of mid-range reading).

The CBW database associates some 8700 persons, 13,000 chapters (biographies), and more than 1200 books of various types published in English 1830-1940 (see developing database at http://cbw.iath.virginia.edu/cbw_db). Our project, however, is neither a textual archive nor a biographical database but an experiment in interpretation using the tools of DH to recognize the conventions of a genre, biography, and the history of gender conventions in a certain social context. Specifically, we want to get at the conditions of nonfiction, which generate multiple versions and cut and paste with relatively little respect for authorship. Could narrative theory of nonfiction be developed through a technique of digital markup that allows us to compare multiple versions of one life and interrelated types of person and text? With Daniel Pitti, Suzanne Keen, postdoctoral Project Manager Rennie Mapp, and teams of graduate assistants, we have developed and deployed a stand-alone XML schema, Biographical Elements and Structure Schema (BESS), in sample archives of digitized collective biographies that include designated individuals (e.g. all collections in our bibliography that include Caroline Herschel, the astronomer).

Briefly, BESS is an XML schema with a controlled vocabulary for narrative elements that appear in a given text:

- **StageofLife:** *before, beginning, middle, culmination, end, after, relative to the lifetime of the biography's subject*
- **EventType** e.g. *illness, persona's*
 - **AgentType** e.g. *mother, unnamed*
 - **Setting:**
 - **Location, e.g. city**
 - **Structure, e.g. school**
 - **Time:** Dates, TimeOfDay, Season
- **PersonaDescription** e.g. *physically daring*
- **Discourse:** e.g. *retrospective, figureOrImage flower*
- **Topos:** e.g. *influence, disgrace*

Each editor in a trained team creates a separate XML file that in effect is an annotated outline, tagging types of elements identified in numbered paragraphs of a TEI file of the biographical narrative (from 3-100+ paragraphs).

```
<event>
  <textUnitRangeReference>
    <start>6</start>
    <end>7</end>
  </textUnitRangeReference>
  <type>marriage, arranged</type>
  <agentType>mother</agentType>
  <agentType>fiance</agentType>
  <agentType>male professional, named</agentType>
</event>
<event>
  <textUnitReference>6</textUnitReference>
  <type>escape</type>
  <type>wedding</type>
  <type>elopement</type>
  <agentType>officer, military</agentType>
  <agentType>lover, male, named</agentType>
  <locationSetting>Ireland</locationSetting>
  <locationSetting>city</locationSetting>
  <dateSingle standardDate="1837-07-23">July 23, 1837</dateSingle>
</event>
```

SAMPLE OF BESS MARKUP OF EVENTS IN COLLECTION a221A D'Auvergne, *Adventuresses* (1927)
Events in bold are common in "Women of the World" books (the sample archive that includes Lola Montez).
Each type of life and type of collection extends the values of the schema.

BESS analysis enables us to compare versions of the same person's life. When we have analyzed all versions in our corpus, we give unique ID numbers to the essential events in all versions (kernels) and the more or less common optional events (satellites, common or rare), and can compare the placement of these in the versions, much as folklorists have charted the variations on the main events of a tale. (Narrative theorists have developed analyses of events in these terms, but not for nonfiction.) BESS analysis reveals differences in narrative technique in books that take different perspectives on women's roles and that select persons of different types. Thus, beyond the literal level of actions (events), we can measurably correlate, for example, the instances of *direct address*, *use of 'we'* alongside not only the *topos* (i.e. situation; underlying scenario) *work as social service* but also the *topos temptation of status or goods*. The conjunction of different elements in these biographies often challenges our own later assumptions about historical women and gender norms. As BESS work is completed, we expand a body of data that for the first time documents the distinctive characteristics of third-person narratives about real people.

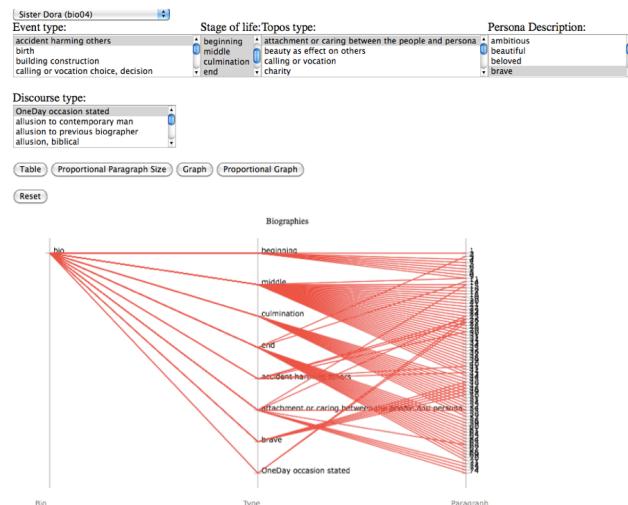


Fig. 1:

Currently working with a web designer, we have a sustainable, accessible database that functions well for team workflow, with parallel display of text and BESS analyses. We plan to develop the visualizations of BESS beyond current designs of tables and graphs.

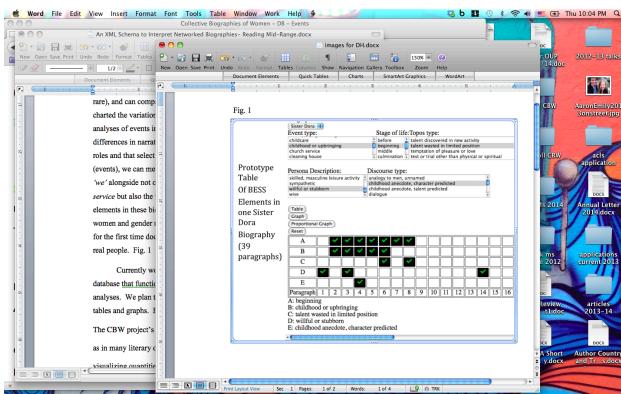


Fig. 2:

The CBW project's BESS "reading" of many narratives is time-consuming and detailed, as in many literary digital projects. As in all DH, we encounter challenges when visualizing quantities of variable data, an issue that this paper merely touches upon. Our aim, instead, is to introduce and make a case for the mid-range approach of BESS, with some reference to other possible approaches.

Many methods of text "reading" en masse might be useful with the CBW books. Broadly, options range from word strings and topic clusters across a large corpus of digitized texts to systematic encoding of all textual features and variants in an onlineß edition of an archive (e.g. Online Froissart; the Rossetti Archive). On the vast end of the scale, we recognize the astounding range of a Google N-gram kind of data capture as well as the precision of some text-mining projects. On the closer focus of the scale, we think human curation is best suited for patterns of narration and ideology, and we begin with books and place them in a context of genre and publishing history. Thus we try to benefit from the precedents of literary editions, and yet CBW is not a project in textual editing. We have no wish to fine-tune exact digital surrogates of these books. The BESS schema and approach to many-versioned biographies within social and historical contexts is designed to moderate between distant and close reading—a comprehensive digital model of variations within a genre and fine textual details.

Many in DH have addressed the question of what to do with a million books. Franco Moretti has espoused distant reading, a term applicable to many kinds of directed and unsupervised queries in big archives. This is understood as opposed to "close reading," the usual literary method (without computational mediation), one text, one person. The findings of singular textual analysis are less appropriate when describing patterns across a genre, especially of nonfiction where there are many versions of the "same" narrative and the author is less important than the protagonist, the representation of a real person. Thus, our method has some affinity for Sharon Marcus and Stephen Best's concept of "surface reading," as advocated in their manifesto to put a stop to the required "critique" or theoretical digging into a text for what it does not say for ideological purposes. Yet we retain a framing conceptual commitment to ideological critique, as we want to know about the changing gender ideology and historical contexts for women's lives. More directly, we are pursuing the kind of "social reading" promoted by Alan Liu, as we hope to extend BESS as a tool available for other projects in digital interpretation of biographical narratives. BESS, interlinked with a database that reveals networks among historical persons and books about them, recruits computation to aggregate the interpretations of readers to parse a genre.

We call for a new metaphor or spatial model for hybrid methods of mid-range digital interpretation, whether using a stand-aside markup like BESS or other approaches. Our editing teams "hover," not at satellite level, but like balloon aerial digital cameras scanning a neighborhood, producing images that can zoom in and out. Such records do much less harm to the person, the individual record, than surveillance or drones. Our schema functions, alternatively, like GPR, "ground-penetrating radar," used to detect buried structures two or three feet into the ground. Although such metaphors for our mid-range reading method with BESS have the inherent

comedy of balloons and robot-like go-carts, we can seriously enhance what we know above or beneath the surface without destroying it—without murdering to dissect. I remind the BESS team that the texts are always still awaiting any method of interpretation, unmangled after we have subjected them to our skewed adaptation for prosopographical purposes. We're not pretending to let machines discover without distortion. Nor are we clinging to the requirement that reading is an individual act—on the contrary. Like reading, biographies trope toward the collective and typological, the mid-range, even in monographic form. Inviting open-access play, we expect to be surprised by the details in the mosaic or wave-like picture from above or below.

References

- Booth, A.** (2004). *How to Make It as a Woman: Collective Biographical History from Victoria to the Present*. Chicago: University of Chicago Press.
- Booth, A.** (2005). "Fighting for Lives in the ODNB, or Taking Prosopography Personally," *Journal of Victorian Culture*, 10: 267-79.
- Booth, A.** "Prosopography." The Encyclopedia of Victorian Literature, ed. Dino Felluga, Pamela Gilbert, and Linda Hughes (Wiley Blackwell), forthcoming.
- Bradley, J. and H. Short.** (2002). "Using Formal Structures to Create Complex Relationships: The Prosopography of the Byzantine Empire—A Case Study." In K. S. B. Keats-Rohan (ed.), *Resourcing Sources Prosopographica et Genealogica*, vol. 7. <http://prosopography.modhist.ox.ac.uk/publications.htm>.
- Cameron, A., ed. (2003). *Fifty Years of Prosopography*. Oxford: Oxford University Press.
- Oldfield, S.** (1999). *Collective Biography of Women in Britain, 1550-1900*. London: Mansell.
- Keats-Rohan, K. S. B.** (2007). "Biography, Identity and Names: Understanding the Pursuit of the Individual in Prosopography." *Prosopography Approaches and Applications A Handbook*. Oxford: Occasional Publications UPR. 139–81. *Prosopographica et Genealogica* 13.
- Stone, L.** (1971). "Prosopography." *Daedalus* 100: 57-9.
- Bauer, J.** *Project Quincy*. <http://projectquincy.rubyforge.org/>
- Brown, S., with Clements, Grundy, et al.** *Orlando: Women's Writing in the British Isles from the Beginnings to the Present*. Cambridge: Cambridge University Press, 2006. <http://orlando.cambridge.org/>
- Clergy of the Church of England Database.** <http://www.theclergydatabase.org.uk/index.html>
- Liu, A.** *Research Oriented Social Environment (RoSE)* <http://rose.english.ucsb.edu/>
- Perdue, S.** *People of the Founding Era*. <http://documentscompass.org/projects/pfe/>
- Pitti, D.** *Social Networks and Archival Context Project (SNAC)*. <http://socialarchive.iath.virginia.edu/>
- Prosopography of the Byzantine World.** <http://www.pbw.kcl.ac.uk/>
- The Prosopography of the Neo-Assyrian Empire** <http://www.helsinki.fi/science/saa/pna.html>

Scholarly primitives revisited: towards a practical taxonomy of digital humanities research activities and objects

Borek, Luise

borek@linglit.tu-darmstadt.de
Technical University of Darmstadt, Germany

Dombrowski, Quinn

quinnd@berkeley.edu
UC Berkeley

Munson, Matthew*mmunson@gcdh.de*

University of Göttingen, Germany

Perkins, Jody*perkintj@miamioh.edu*

Miami University, United States of America

Schöch, Christof*christof.schoech@uni-wuerzburg.de*

University of Würzburg, Germany

1. Introduction

Today we have more information at our fingertips than at any other time in human history. The problem is no longer finding information, the problem is being overwhelmed with the amount of information. This is no different in the realm of the digital humanities. Information on people, projects, resources, methods, and tools exists in quantity everywhere we look, and yet we still have difficulty finding what we need. This paper will describe a transatlantic effort on the part of DiRT in the United States and Dariah in Europe to construct a taxonomy of scholarly methods, that can be used not only to organize single collections of DH information and resources but also to allow these collections to interface with each other, creating a web of linked data that can be effectively searched for information across distributed collections. DiRT and Dariah are not trying to impose a restrictive, monolithic scheme on DH; rather, our goal is to construct a lightweight, basic taxonomy of higher order goals and first-order methods that can be easily expanded in all directions by linking lower order techniques to multiple goals and/or methods to create machine-readable paths among the various resources. In building this taxonomy, we heavily rely on input and feedback from the digital humanities community. Still, at least for the intended use cases, we believe a stable taxonomy has advantages over more open, folksonomy-based solutions.

The taxonomy as it exists now is based upon three primary sources: 1) the arts-humanities.net taxonomy of tools of DH projects, tools, centers, and other resources, especially as it has been expanded by *digital.humanities@oxford* in the UK and DRAPler in Ireland; 2) the DiRT collection of digital research tools, re-launched under Project Bamboo in the US but now continuing on after the end of that project; and 3) the Dariah 'Doing Digital Humanities' Zotero bibliography of literature on all facets of DH. These resources were studied and distilled into their essential parts, producing a simplified taxonomy of two levels: 8 top-level goals that are broadly based on the steps of the scholarly research process and a number of general methods under these goals that are typically used by scholars to achieve these research goals. The updating of the taxonomy and the definition of the types of relationships to be described in the resulting ontology will be carried out by a joint working group in the Dariah-EU and the NeDiMAH projects in Europe, which will conduct large scale desk and field research into scholarly practice to determine how best to describe the relationships between and among the goals, methods, and techniques of scholarly practice. The future expansion of this organizational system will not be as a hierarchical taxonomy but, instead, as a linked ontology as lower-level techniques are attached to one or more methods, linking all the existing entities in the ontology together. The projects and collections that use this schema will play an important role here: as resources are added to these collections and linked to the taxonomy, the resulting ontology will grow in complexity. This complexity will be more help than hindrance precisely because it will be a machine-actionable complexity. Computers will traverse this web of relationships for us, only bringing back results that are closely related to our needs.

This may seem excessively optimistic, but this paper will support these claims by describing three very different types of resources that have used and expanded the taxonomy not only to improve the findability within their own collections but, more importantly, to link to each other in a machine-actionable way. These resources are the DiRT directory of digital tools, the Dariah 'Doing Digital Humanities' bibliography, and the

Dariah-DE service-oriented project portal. A brief description of each of these collections and how they will profit from this taxonomy/ontology follows.

2. DiRT

DiRT (Digital Research Tools, <http://dirt.projectbamboo.org>) is a longstanding US-based directory for scholars interested in digital tools, which provides basic information about software that can facilitate the research process at different stages. The classification of tools by category has always been fundamental to DiRT: in its earlier incarnation as a wiki, wiki pages each corresponded to a category of tools, and the tools were presented in a list on the page. In 2011, DiRT was rebuilt using the Drupal content management system, which allowed information about each tool to be stored in a structured manner that enables faceted search and browsing. While users can now create complex queries on DiRT (e.g. using operating system and price to narrow their results), tool categories remain the primary way of navigating the site.

With support from the Andrew W. Mellon Foundation, DiRT is currently undergoing a new phase of development, with the goal of making information about digital tools available outside the DiRT directory itself. Since its inception, DiRT has used its own ad-hoc list of categories. All tools must belong to at least one category, though these categories can be supplemented with user-generated tags. The shortcomings of DiRT's categories list can be illustrated through the example of OCR tools-- some are classified as "transcription", others as "conversion", and while neither is ideal, both are a reasonable approximation given the other options. Replacing DiRT's former categories with the taxonomy will improve the consistency and quality of the data, and also provide a shared facet that can connect DiRT's tool data with information provided by other projects, once DiRT's contents are made available using RDF.

3. 'Doing Digital Humanities' bibliography

Another resource directly connected to the taxonomy is Dariah-DE's 'Doing Digital Humanities' bibliography. The bibliography can be accessed on Zotero (www.zotero.org/groups/113737) or on the Dariah-DE portal (<https://de.dariah.eu/bibliography>). Like DiRT, the bibliography is one of the seed activities for the taxonomy at the same time as being one of the already defined use cases, representing the application domain of making medium-sized collections of bibliographic references discoverable. This Zotero-based bibliography offers suggestions for introductory readings as well as more in-depth coverage of research literature in various areas of digital research, teaching and infrastructure planning in and for the humanities. The bibliography is carefully curated collaboratively, is freely accessible, currently has around 800 entries, and is being updated continuously.

Right now, the bibliography is already divided into thematic collections based on the "goals" defined in the taxonomy. Each collection, hence, covers one prototypical aspect or goal of the research process in the humanities as it is practiced with digital tools, methods and data. In addition, all entries in the bibliography are discoverable through keywords covering, on the one hand, typical research methods and activities in the humanities, and on the other hand, a wide range of objects of research. The current closed list of keyword represents an early draft version of the taxonomy described here.

Once a first stable version of the taxonomy is available, the bibliography's keyword implementation will be updated. Sharing a keyword system with other projects will make it easier for users to find related resources. And the public documentation of the taxonomy, including concise scope notes for all methods and techniques, will make the bibliography's keyword-based search more transparent and increase its usability.

4. Dariah-DE portal

A third use case aims to examine the taxonomy as a functional structure for Dariah-DE's service-oriented website, the Dariah-DE-Portal. Launched in a first version in May 2013, it will receive a makeover in the early stage of the upcoming German Dariah II project scheduled for March 2014 that is based on the taxonomy.

The website is designed to offer a wide range of services concerning Digital Humanities in Germany and addresses both researchers who already work digitally and those seeking information or advice. The services provided are as heterogeneous as the DH landscape. They cover informational aspects on specific research projects, information on DH Centers, Bachelor/Master Programmes and tools as well as their documentation, tutorials and teaching materials. Services offered by Dariah-DE (like the embedded bibliography mentioned above, the Dariah-DE Working Papers, or hosting services and a developer's portal) are complemented by external resources like blogs and a DH-calendar (a cooperation with calenda.org being currently on its way).

The variety of this content leads to multi-purpose requirements that enable a flexible access to information relevant to individual users. This use case meets that challenge by implementing the taxonomy in RDF, thus interlinking content and making it multi-purpose. In that way, the taxonomy will function as a 'meta-service' that meets the interests of an active and interlinked community, that visualizes Digital Humanities and promotes its results.

5. Conclusion

The purpose of this talk is not to convince the audience that we in DiRT and Dariah have all the right answers. Instead, it is to continue a conversation about the importance of ontologies for managing the over-abundance of DH information, present our own work on this problem and our approach to gathering and incorporating community feedback, in hopes of spurring further work in this area.

References

- Anderson, Sheila; Tobias Blanke; Stuart Dunn.** (2010). *Methodological Commons: Arts and Humanities E-Science Fundamentals*. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 368, no. 1925 (2010): 3779–3796. <http://rsta.royalsocietypublishing.org/content/368/1925/3779.abstract>.
- Benardou, Agiatis, Panos Constantopoulos, Costis Dallas, and Dimitris Gavriliis.** *Understanding the Information Requirements of Arts and Humanities Scholarship*. International Journal of Digital Curation 5, no. 1 (June 22, 2010): 18–33. doi:10.2218/ijdc.v5i1.141.
- Borgman, Christine** (2010). *Scholarship in the Digital Age : Information, Infrastructure, and the Internet*. Cambridge: MIT Press.
- CLIR (Commission on Cyberinfrastructure for the Humanities and Social Sciences)**. (2006). *Our Cultural Commonwealth: The Report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences*. New York: American Council of Learned Societies, 2006.
- Gasteiner, Martin, and Peter Haber**, eds. (2010). *Digitale Arbeitstechniken für die Geistes- und Kulturwissenschaften*. Vienna: UTB. <http://www.utb-shop.de/digitale-arbeitstechniken.html>.
- Reiche, Ruth; Rainer Becker; Michael Bender; Matthew Munson; Stefan Schmunk; Christof Schöch.** (2014). *Verfahren der Digital Humanities in den Geistes- und Kulturwissenschaften*. Dariah-DE Working Papers Nr. 4. Göttingen: Dariah-DE (to appear). Preprint: https://dev2.dariah.eu/wiki/download/attachments/2295542/M223_DH-Verfahren.pdf.
- Siemens, Ray; John Unsworth; Susan Schreibman**, eds. (2004). *A Companion to Digital Humanities*. Hardcover. Oxford: Blackwell. <http://www.digitalhumanities.org/companion/>.

Unsworth, John. (2000). *Scholarly Primitives: What Methods Do Humanities Researchers Have in Common, and How Might Our Tools Reflect This?* London: King's College London. <http://www3.isrl.illinois.edu/~unsworth/Kings.5-00/primitives.html>.

Single Page Apps for Humanists: A Case Study using the Perseus Richmond Times Corpus

Borg, Trevor

Loyola University Chicago, Center for Textual Studies and Digital Humanities

Thiruvathukal, George Kuriakose

Loyola University Chicago, Department of Computer Science ; Loyola University Chicago, Center for Textual Studies and Digital Humanities

TEI is good at what it does: static documents rendered in glorious detail. But TEI is old. Its age doesn't make TEI irrelevant, but it's important to be conscious of how the way we weave the fabric of the web has changed since TEI was conceived in 1994, and reevaluate some of our assumptions about its use. In this early work, we are exploring this rethinking as part of a larger study within the center on general methods for isolating the complexity frequently associated with XML-based frameworks.

The *Richmond Times Dispatch* corpus of TEI-encoded newspapers comprises the Confederate newspaper's Civil War run, 1860 — 1865. It is compelling both in terms of organization and content and amounts to a comprehensive textual index. In addition to the historical allure of its content, the formal properties of the digitized documents made available through the Perseus Collection make the *Dispatch* an extraordinary raw material for building a rich interactive visual experience that augments the *textual* one.

The *Dispatch* is not in need of a new home; the Perseus Collection hosts a perfectly *functional* version and uses best practices for data encoding and organization. Rather than strive to be the primary resource for the source material, our project uses the source material to explore recent patterns in web development as well as alternative, more visually compelling ways to interact with XML corpora in a web application. The goal is to produce a powerful reading environment that is tailored to its source material to an extent that the generalized project of the Perseus Collection can afford.

With respect to its implementation, our project fits the genre of a 'single page application' (SPA). This project demonstrates best practices for implementing this type of software project using a particular suite of tools; as an open source example of an SPA that is considerably more complex than the usual teaching examples for this kind of thing, we hope that our implementation will be useful to other people who are considering using the same tools, and especially to humanists interested in presenting TEI-encoded documents.

The SPA is *du juor*. We prefer beautiful URLs and smooth transitions. We are less fond of all these big lists cluttering our sidebars and clunky arrays of checkboxes. We don't expect websites to always be inert collections of documents. We want to be able to control the connective tissues. The web development community has responded to our current expectations with tools to suite them.

The client-side MVC (Model-View-Controller) libraries that have recently emerged have reached a high level of maturity. A client-side MVC library codifies conventional solutions to the generic problems posed by web traffic. It provides semantics for describing the interaction layer between data and presentation. The codification of conventions that MVC libraries manifest is exciting. It deeply simplifies matters for those who want to make interactive documents. Humanists who have a grasp of the language and concepts involved will be that much better able to articulate and realize project architectures that delight the contemporary reader.

For instance, our application is built around a *client-side router*. The router formalizes protocols for state transitions that allow for timely and efficient request management. We rely heavily on the concept of the *run loop*, which exposes powerful document management techniques and is tightly linked with a client-driven *templating engine*. We are able to achieve a remarkably clean separation of concerns in a highly condensed space by exploiting the conventional roles organized and implemented by these libraries. And by shifting our application's emphasis to the client, we have constant access to a unified programming environment, limiting the context-switching required when developing different parts of the application.

In addition to our project's strong client-focused application architecture, we also demonstrate a data architecture solution to the problems posed by the corpus' rich TEI markup. To expose the facets embedded in the source XML, the implementation transforms the deeply nested structure inherent into flat relational representations that can be searched efficiently. Furthermore our project demonstrates a novel, pythonic approach to transforming the source XML to browser-ready HTML that is particularly amenable to the constraints of an SPA.

XSLT wasn't very well suited to our TEI transformation problem. One of the key UI features of our application was the ability to discover and search for special entities such as people and places in the text. By implementing a custom transformer in python, we had the flexibility to both translate the TEI tag names into valid HTML versions and retain the original TEI tag names and attributes as attributes on the HTML element.

In addition to serving content thus transformed as needed, the role of the server in our application is limited to various precomputation and preprocessing tasks that only need to be run whenever the source material changes--a process that is fully automated with Unix batch processing (via cron) in the cloud. Users never notice. Research projects are often quagmired in a chaotic sprawl of one-off scripts; we demonstrate a coherent architectural pattern for orchestrating these preprocessors.

Sometimes the affordances of an SPA make it worthwhile to depart from the original document's presentation. Content on the web wants a different kind of exposure than a stack of newspapers. You want to be able to find things quickly. You want to be able to highlight and hyperlink, associate and drill down. Once you've computed a graph of your stack of newspapers, now you can move laterally, staying in the same section and moving from date to date, just as easily as you could stay on the same date and move top-to-bottom through the articles. We demonstrate a novel, minimalistic navigation scheme for the *Dispatch*.

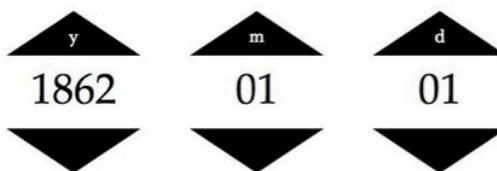
If you're taking full advantage of a Javascript environment to render your XML content, you can use modern libraries to plug in visualizations with simplicity, and furthermore to turn these visualizations into interactive filters for a very powerful browsing experience. Using a cluster of technologies surrounding Mike Bostock's work, we demonstrate how to integrate a visualization library into an SPA.

And yet we believe that datafication shouldn't overwhelm the content. You want to be discrete about placing your controls lest you scare the casual user, but they should be powerful. Live feedback from search inputs has come to be a common expectation for user interfaces and the SPA environment makes it easy to architect that. We show one effective way to make your XML live-searchable.

We close with just a few screenshots of the work in progress. It is important to note that this interface is being further refined based on new work Trevor is doing in his new role as a front-end software developer.

THE RICHMOND TIMES DISPATCH

Daily news from the Confederacy's most permanent capital, Nov. 1860 - Dec. 1865



news

telegraphic

local

morning

Fig. 1: An early version of the splash page, presenting user interface controls for the issue date, section, and subsection. Selecting a subsection would reveal the list of headlines it contains. The date selectors reload the issue content asynchronously, without reloading the page.

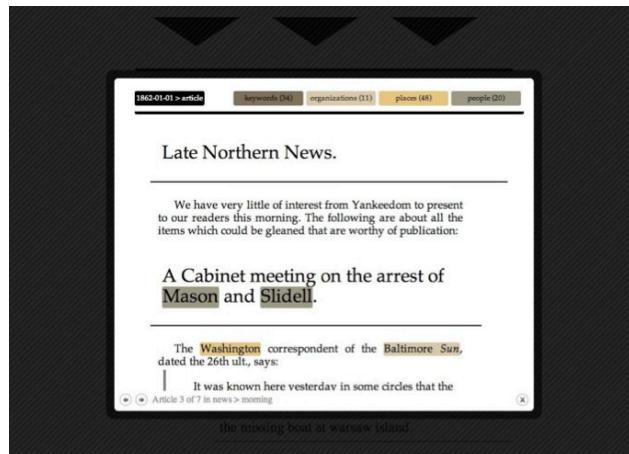


Fig. 2: The early version of the article reader, presenting the text in a modal context. The summary of facets across the top act as toggles for corresponding highlights in the text. You can page through the section content using the controls at the bottom of the modal.

names of Banks.	Capital	specie	circulation.	Deposits.	Discounts
Bank of Virginia	2,651,250.00	431,830.77	1,230,912.83	2,138,374.74	4,090,304.10
Bank of Berkeley	100,000.00	13,889.51	60,160.00	33,678.04	61,996

Fig. 3: This screen capture shows the direction taken in the latest development. The URL bar demonstrates stateful client-side routing. The content selection controls have been flattened into a trio of type-ahead controls with pagination buttons for navigating forward and backwards both in the document structure and across documents.

This stuff is fun. The tools are a joy to use. The free/open source community behind it is excellent and innovative.

We want to see more humanists building applications, and moving away from consuming and rather heavyweight content management systems such as Drupal. Based on our experience, humanists can learn the tools and frameworks quickly with excellent results to boot. We hope that our implementation of the *Dispatch* will set a strong example for our (and others') future DH projects.

References

We note that we are proponents of using XML, especially for its originally intended purposes of self-describing data interchange, which remains tremendously valuable in developing type safe RESTful web services.

George K. Thiruvathukal is leading a separate and parallel effort to develop the Standoff Markup text editor, standoffmarkup.org, which is aimed at simplifying the encoding and maintenance of XML texts (without exposing tree-oriented abstractions). This is where we started exploring the use of SPA when it comes to building DH-facing tools in general.

Richmond Times Collection, www.perseus.tufts.edu/hopper/collection?collection=Perseus:collection:RichTimes

Single Page Applications original conception, code.google.com/p/trimpah/wik/SinglPageApplications

Conventionally, a to-do list. See, for instance, todomvc.com.

The Model-View-Controller design pattern (and paradigm) was introduced as part of the Xerox PARC Alto computer, which used the Smalltalk programming language. An excellent historical read about this paradigm can be found at heim.ifi.uio.no/~trygver/themes/mvc/mvc-index.html.

We use Ember.js: emberjs.com. Our technical term for "high level of maturity" is "rock" but we eschew this Americanism for the purpose of a conference paper submission.

We rely throughout on the wonderful lxml library for Python: lxml.de.

Our present implementation is deployed on Heroku, an agile and scalable framework for deploying apps like ours. The overall project is moving to Linode (a dedicated Linux-based cloud hosting provider).

In this we rely on the large-scale application planning features offered by the Flask web framework for python: flask.pocoo.org.

Author of d3js.org among other things.

We have found dc.js (nickqizhu.github.io/dc.js) to be a perfect storm of visualization functionality.

We use the well-known elasticsearch library (www.elasticsearch.org) to achieve an effect like Google's live search results.

A top-down approach to the design of components for the philological domain

Boschetti, Federico

federico.boschetti@ilc.cnr.it
ILC-CNR of Pisa (ITALY)

Del Grosso, Angelo Mario

angelo.delgross@ilc.cnr.it
ILC-CNR of Pisa (ITALY)

Khan, Anas Fahad

fahad.khan@ilc.cnr.it
ILC-CNR of Pisa (ITALY)

Lamé, Marion

marion.lame@ilc.cnr.it
ILC-CNR of Pisa (ITALY)

Nahli, Ouafae

ouafae.nahli@ilc.cnr.it
ILC-CNR of Pisa (ITALY)

Introduction

This paper focuses on the methodology applied to the development of components in the domain of collaborative philology in the Memorata Poetis Project. This initiative, led by the University of Venice, coordinates eight units sharing the same cyber-infrastructure and is co-funded by the Italian Ministry of Instruction, University and Research (PRIN 2010/11).

The project aims to study the multilingual intertextuality between epigraphic texts and literary epigrams, the transmission of themes, motives, etc. between different communicative situations (epigraphic versus literary) and different civilisations (Greek, Latin and Italian). As a control group, we analyse a corpus of epigraphic and literary texts in Arabic which do not belong to the same tradition as the others. The study of intertextuality affects both the reconstruction of the text (*constitutio textus*), by providing variants from the indirect tradition, and its interpretation (*interpretatio*), by widening the contexts in which the text has been reused.

Methodology

By following a top-down approach the article will discuss the following three aspects of the general design of the components developed by the Institute for Computational Linguistics of the National Research Council (ILC-CNR) in Pisa, which will be integrated into the shared infrastructure managed by the Venetian working unit of the project.

Firstly, we will introduce the ongoing modelling of the philological domain from a formal point of view. Secondly, we will discuss engineering methods for the analysis of the required components. Finally, we will describe the application of the aforementioned methodology to the specific part of the project developed in Pisa.

Computational philology has so far focused on the formalisation of only some aspects of the philological domain, such as stemmatics, derived from the Lachmannian methodology¹, but it is necessary to take into account the formalisation of other aspects essential to understanding the history of the tradition² as well as the relation that a text has with its text bearing object (TBO), reusing non-textual annotation tools³. Thus, any proposed formal models should reflect a representative range of philological methods and practices.

Whereas stochastic theories and processes borrowed from computational linguistics have been successfully employed in computational philology, formal models based on selected logical axioms specific to the philological domain, have not been sufficiently developed⁴. In this aspect of our work, our attention is addressed to an overall class of problems rather than just a single project. The ultimate goal is to model how various kinds of philological data serve as evidence for the construction of dynamic critical editions and critical commentaries. As another outcome, these logical models might result in the development of an extensive domain ontology and subdomain ontologies.

An example should illustrate the benefits that such a process of formalisation could have in the development of software tools for projects in the philological domain such as Memorata Poetis. An analyst designing software for a project that must deal with textual variance due to the existence of several diverging manuscripts of the same work, can afford to focus on creating tools to handle different chunks of text starting at the same textual position, as per his design specifications, while neglecting to deal with the issue of multiple syntactic interpretations in ambiguous sentences. A different project requiring such an extension to the original software in order to record concurrent syntactic analyses suggested by different scholars in commentaries will have to incorporate a comprehensive process of refactoring, instead of a simple development that extends the functionalities of the software developed in the previous project.

Much work in computational philology in the last few decades has been driven by the idea that the design and development of a digital platform for text criticism can be carried out by simply

transferring and customizing many of the tools that have been developed in the field of computational linguistics for studying modern languages^{5, 6, 7}. However we think that it is necessary to develop a different line of research in which the tradition of philological studies can advance into the digital era without relying on such a simplistic view of the relation of such work with computational linguistics.

The development of software components for the philological domain at ILC-CNR is based on the agile paradigm of software development: we mix a top-down with a bottom-up approach, which requires a continual improvement in design and implementation.

Ongoing Results

The library of core components under development is structured into the following packages: philological content management, TBO management, editing, management of layers of analysis, relations (linked data) management, indexing, search, view.

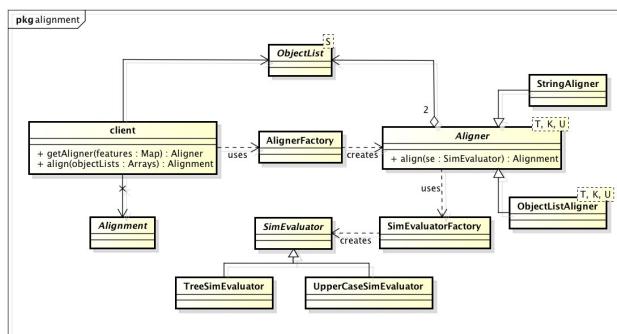


Fig. 1: Class Diagram of the Aligner Component

Philological entities can be either represented as linear or non-linear structures; in the latter case we have the choice of representing textual variants as graphs⁸ or in other ways (e.g. as a swarm of variants). The choice is determined on the basis of the best trade-off between fast access, representation of variable granularity, etc. The strategy for the actual representation of texts with variants will be implemented in the extended classes of the abstract PhilologicalEntity class, which provides methods to set and get the textual variants.

TBO components deal with information related to the epigraphic device, in our case a small subset of the epigraphs. These components manage the multidimensional models (e.g. 3D) and any other relevant information related to the epigraphic situation. Epigraphy, as a specific communication process of written text, gives complementary examples of the scientific and digital requirements for a global approach of the TBO. By focusing, among many other complex aspects, on writing and context, epigraphy concerns itself with entangled information from the process of communication that computational linguistic processes only partially take into account. This is necessary to the overall scientific interpretation and understanding of any text.

Editing components manage the creation, reading, updating and deletion of the data stored in the system, preserving the integrity of the data, tracking multiple versions of the information, etc. The following types of objects are affected by editing: texts with variants, automated analyses described below (in order to manually review them), data entries for free annotations (such as commentaries) and structured annotations (such as the tagging of themes and motives and semantic analyses, according to the SIMPLE methodology⁹).

Components related to linguistic and stylistic automated analyses both implement cutting-edge algorithms for lemmatization and pos-tagging¹⁰ as well as embedding tools developed in the Perseus project like Morpheus. Components for metrical analysis¹¹ and¹², individuation of named entities, etc. are pluggable extensions.

Here it is interesting to note that adapting computational models developed for Western languages could result in the

loss of information regarding innate characteristics of different and more remote languages as pointed out in recent projects such as Sharing Ancient Wisdoms (SAWS-KCL). For instance the word analyses made by Buckwalter's morphological engine are not marked according to Arabic grammar but according to their translation in English¹³. For example, the word biHaq-i is analysed as a preposition and this is incorrect. The words commonly used to translate biHaq-i in English, e.g., "against", are indeed prepositions, but in Arabic grammar, biHaq-i is composed of the concatenation of three parts: (1) bi=PREP + (2) Haq~NOUN+ (3) i=CASE_DEF_GEN. For these reasons, we have brought about improvements to the current morphological analyzers which allow detailed analyses respecting the grammar and granularity of Arabic¹⁴. Linked data components will be developed in order to handle the overall relations between the entities involved in the system through an identification scheme (e.g. RDF). The linking is done at different levels of granularity and between different types of objects. For example, a philological entity can be linked to another philological entity and a character can be linked to the related box in its three-dimensional model. Indexing components will create and handle data structures necessary to efficiently access stored resources. Search components, devoted to information retrieval, will combine the data indexed in the persistence unit and exploit a large number of query techniques for accessing databases (xpath, sql, sparql, etc). View components will take into account the data structures that represent content combined with multiple levels of analysis. The interaction between the user and the system through the graphical interface (user experience) must be suitable for philologists and their specific needs, avoiding limitations due to the adaptation of the user experience of different domains.

Italian Selected Text			Arabic Selected Text		
Form	Lemma	Analisi	Form	Lemma	Analisi
Questa	questo	PDI+num+s+gen+f	هذا	هذا	h/*+A=DEM_PRON_MS+
è	essere	V+num+1+par+3+mod+= ts =p	فِي	فِي	qabor+NOUN+vaCASE_DEF_NOM+
la	il	RD+num+s+gen+f	الـ	الـ	Al+DET+Babot+NOUN+=iCASE_DEF_GEN+
sepoltura	sepoltura	S+num+1+gen+f	القبر	القبر	Al+DET+f+dqlyr+AD)+=iCASE_DEF_GEN+
di		EA+num+s+gen+m			
cervo	servo	S+num+1+gen+m	لـ	لـ	< I>Y+PREP+

El-Bi intrò agli scavi di Alessio, dove fu fatto un
ILC-CNR 2013

Fig. 2: Web Interface showing the text of an Arabic epigraph aligned with its Italian translation and related morphological analysis

Conclusion

In conclusion, our approach tries to model the principal entities, their relations and their behaviour in the domain of philology at a high level of abstraction and, consequently, we derive a framework that is not based on the requirements of a specific project, but that derives from the logical modelling of the domain. Eventually, the actual software components developed according to the framework will be used for a collaborative project that combines multiple levels of analyses and annotations, in order to enrich the traditional methods applied by philologist to study intertextuality. Applications developed with the CoPhi components are made available here: <<http://cophilab.eu>>.

References

- Roos, T. and Heikkila, T. (2009). *Evaluating methods for computer-assisted stemmatology using artificial benchmark data sets*, Literary and Linguistic Computing, 24: 471-433.
- Bozzi, A. (2004). *Postfazione a Zampolli Antonio, Filologia e informatica: le origini della filologia computazionale*, Euphrosyne, n.s. 32: 21-24.
- Soler, F., Torres, J. C., León, A. J. and Luzón, M. V. (2013). *Design of Cultural Heritage Information Systems based on Information Layers*, ACM Journal on Computing and Cultural Heritage. In press.

4. **Endriss, U.** (2011). *Logic and social choice theory*. In Gupta, A. and van Benthem, J. (eds.), Logic and Philosophy Today. London: College Publications. staff.science.uva.nl/~ulle/pubs/files/EndrissLPT2011.pdf (accessed 7 March 2014)
5. **Bamman, D. and Crane, G.** (2009). *Computational Linguistics and Classical Lexicography*, Digital Humanities Quarterly, 3. www.digitalhumanities.org/dhq/vol/3/1/000033.html (accessed 7 March 2014).
6. **Robinson P.** (2004). *Where We Are with Electronic Scholarly Editions, and Where We Want to Be*, Jahrbuch für Computerphilologie 5: 123-143. computerphilologie.uni-muenchen.de/ejournal.html (accessed 7 March 2014).
7. **Orlandi, T.** (2010). *Informatica testuale - teoria e prassi*, Roma: Laterza Editori.
8. **Schmidt, D. and Colomb, R.** (2009). *A data structure for representing multi-version texts online*, International Journal of Human-Computer Studies, 67(6): 497-514.
9. **Lenci, A., Calzolari, N. and Zampolli, A.** (2003). SIMPLE: Plurilingual Semantic Lexicons for Natural Language Processing, Linguistica Computazionale 16-17: 323-352.
10. **Bamman, D. and Crane, G.**, (2011). *The Ancient Greek and Latin Dependency Treebanks*. In Sporleder, C., van den Bosch, A. and Zervanou, K. (eds.), Language Technology for Cultural Heritage. Berlin: Springer Verlag, pp.79-98.
11. **Pavese, C. O. and Boschetti, F.** (2004). *A Complete Formular Analysis of the Homeric Poems*, Amsterdam: Hakkert.
12. **Fusi, D.** (2004). *Fra metrica e linguistica: per la contestualizzazione di alcune leggi esametriche*. In Di Lorenzo, E. (ed.), L'esametro greco e latino: analisi, problemi e prospettive - Atti del convegno di Fisciano 28-29 maggio 2002. Napoli: Guida Editore, pp.33-63.
13. **Zemirli, Z. and Elhadj, Y. O. M.** (2012). *Morphar+: an Arabic morphosyntactic analyzer*, Proceedings of International Conference on Advances in Computing, Communications and Informatics (ICACCI), Chennai, India, 3-5 August 2012, pp. 816-823.
14. **Hajder S. R.** (2011). *Adapting Standard Open-Source Resources To Tagging A Morphologically Rich Language: A Case Study With Arabic*, Proceedings of the Student Research Workshop associated with RANLP 2011. Hissar, Bulgaria, pp. 127-132.

Exploring a model for the semantics of Medieval Legal Charters

Bradley, John

john.bradley@kcl.ac.uk

Department of Digital Humanities, King's College London, United Kingdom

Rio, Alice

Department of History, King's College London, United Kingdom

Hammond, Matthew

Department of History, University of Glasgow, United Kingdom

Broun, Daivit

Department of History, University of Glasgow, United Kingdom

Historical legal charters can provide an important window into the workings of a society, and because of their legal significance they often survive as evidence of a society when there is little other documentary evidence available. Of course, original medieval charters have always been difficult to access, and for this reason work was undertaken even in the 19th century to prepare printed editions of them. Nowadays they are published over the WWW, since space is unlimited and access for potential users is quick and easy. See, for example, the efforts of the Institute of Historical Research in its *British History Online project* which has been transcribing printed editions of collections and making them available over the WWW (e.g. Marwick 1894), and the *monasterium.net* (Mom 2013) from which, as its website says, "Original documents [are] available world wide 24 hours a day, independent of the researcher's physical location."

Once the texts have been digitised, the question arises about what can be done with them. One strand for this is exemplified in the work of the ChartEx project (ChartEx 2013) which has been funded by the *Digging into Data Challenge* and is developing "new ways of exploring the full text content of digital historical records" through the use of natural language processing and data mining techniques to extract automatically information about people, events and places and find relationships between them. ChartEx, then, by its use of advanced text processing and analysis techniques follows one of the long traditions of the digital humanities. The other long standing approach to dealing with text – markup – is also active with the work of the *Charters Encoding Initiative* project (CEI 2009) which has developed a TEI-derived markup scheme suitable for charter texts. CEI takes, perhaps not surprisingly, primarily a diplomatic approach. Georg Vogeler, writing about the CEI, claims that medieval European charters "reflect contemporary attitudes and mindsets as regards legal and representation issues" and "are tools of diplomatic criticism" (Vogeler 2005, p 276). He further claims that through CEI markup one has a "platform for seeing the European Middle Ages as they are reflected in their charters" (Vogeler 2005, p 279).

DDH at King's College London has been involved in several projects that have used charters as their subject matter, but these projects have not taken either the text analysis or markup approaches. Charters were a substantial component of the *Prosopography of Anglo-Saxon England* (PASE 2010). Two other more recent projects have made charters their primary focus, even more than PASE. The *Paradox of Medieval Scotland* project (PoMS 2010) (extended more recently in the *People of Medieval Scotland* (PoMS 2013)) represents more than 6000 charters and claims to be "the first online prosopographical database to comprise exclusively and exhaustively the corpus of administrative documents of a European kingdom in the central middle ages." (Hammond 2013, p 7). More recently, and still in development, the *Making of Charlemagne's Europe* project (Charlemagne 2013) is "a database of prosopographical and socio-economic data found in the more than four thousand legal documents surviving from Charlemagne's reign." An important difference between these projects and the methodologies implied by projects such as Chartex or CEI is that none of PASE, PoMS or Charlemagne actually represent the texts themselves. Instead all three projects take a structured data approach – developing a formal model in highly structured data that aims to express some aspects of what the charters are talking about. The very nature of charter texts is that they provide what their creators thought of as a formal approach to the dealing with their property. The models used for these three projects try to capture explicitly formal structures that the charter's creators seemingly had in their mind when they created them. Although studies of the charter documents such as that carried out in diplomatics provides one of the bases for the work here, as Hammond says about PoMS: "A clear distinction has been drawn [in PoMS] between the form of the document, based on its diplomatic, and the events which are transactions ... in the document" (Hammond 2013, p 11). What can be said about the charters in these projects when the charter text itself is not present? It turns out that the absence of the text enables a representation of the material about a charter that somehow liberates it from being grounded too exclusively in the words of the text itself.

All three of these projects grew out of the thinking about structure that is represented by DDH's factoid approach to prosopography, described in Bradley and Short 2005 and extended into the world of the Semantic Web in Pasin and Bradley 2013. The virtue of the factoid approach is that it provides a source-grounded place for our researchers to say many different kinds of things about their documents. It allowed researchers to record not only the main things the charter was created to talk about – the arrangements about property – but also the wealth of further information that they capture: relationships between people (kinship and otherwise), the titles they held, other events they were involved in, etc. The factoid model also allows for the roles of people and associated other objects mentioned in the charters to be specified and

for the way in which their name was recorded in a particular spot in the charter text to be recorded as well. Furthermore – and most relevant here – it supported a richer understanding of a charter's complex set of associated events. In PASE, for example, we were able to formally separate the event of the charter signing, with its perhaps specific historical significance, from the activities – usually transactions on property – that the charter talked about. Multiple exchanges in a single charter could be captured as separate transaction-like event factoids.

In the end the factoid approach to the charters encouraged us to also think about how to formalise other aspects of what our charters represented. People, for example, turn up not only as agents in the charters but sometimes as possessions. Possessions are complex entities in charters: often they are pieces of land or institutions sitting on land, but charters also show people possessing rights of many kinds, such as the right to run a fair, to collect taxes, to move away from a piece of property, to celebrate divine service, etc. Furthermore, what was actually being given in relationship to a possession could be complex too. PoMS researcher John Reuben Davies developed a scheme to clarify and enrich the classification of transactions presented in charters, going beyond simple classifications such as 'grants' and 'confirmations' to include 'renewals' and 'successions', among many others. *Charlemagne* has been exploring the terms and conditions attached to the exchange of possessions as described in their charters, and has also been exploring the complex way in which historical places are presented in their charter documents – not only mapping them, where possible, to modern places, but attempting to represent the complex and often obscure connections between place and historical region that their charters describe. In the end, the rich formal structure that underpins PoMS and *Charlemagne* represents some of the subtlety of the legal understanding that was emerging in Charlemagne's time and in the days of Medieval Scotland. It exposes some part of what the charter creators were trying to achieve in their efforts to formalise their agreements between themselves about their possessions.

In this presentation we will introduce some of the structure that extends the factoid model and represents several of the complex processes that these charter documents were attempting to formalise. The resultant models attempt to structure aspects of these documents than have not been tackled before and tries to find a point that both enriches our understanding through formalising them, and avoids ignoring, through excessive formalisation, the ambiguity and vagueness of the emerging legal process that happened both in the time of Charlemagne and Medieval Scotland. By developing a formal model for our understanding of the framework in which the charters operated, we believe that we complement the text based approaches of both *Chartex* and the *CEI*.

Finally, our work is, as the famous historian E.H. Carr said, an attempt to recognise "History [as] a process". With its extension of the charter documents into the historical world of people, places and possessions, it attempts to recognise that "you cannot isolate a bit of the process and study it on its own" (Evans 2001). Our models make explicit a set of views and assumptions by our researchers about these medieval worlds, and in their formality and clarity make it more possible to, in the words of the historian Richard J. Evans, "subordinate them to the intractabilities of the material with which they are working, and enable readers to study [our] work critically by making these views and assumptions explicit". In this presentation we hope to encourage this kind of dialogue.

References

- Bradley, John and Harold Short** (2005). "Texts into databases: the Evolving Field of New-style Prosopography" in Literary and Linguistic Computing Vol. 20 Suppl. 1:3-24.
- CEI** (2009). *CEI – Charters Encoding Initiative*. Online at <http://www.cei.lmu.de/index.php> (Accessed 30 October, 2013).
- Charlemagne** (2013). *The Making of Charlemagne's Europe*. Online at <http://www.charlemagneseurope.ac.uk/> (Accessed 30 October, 2013).
- ChartEx** (2013). *ChartEx*. Online at <http://www.chartex.org/index.html> (Accessed 30 October, 2013).
- Evans, Richard J.** (2001). "The Two Faces of E.H. Carr". In What is History?. Website published by Institute of Historical Research. Online at <http://www.history.ac.uk/ihr/Focus/Whatishistory/article.html> (Accessed 30 October, 2013).
- Hammond, Matthew** (2013). "Introduction: The Paradox of Medieval Scotland". In Matthew Hammond (ed). New Perspectives on Medieval Scotland 1093-1286. Woodbridge UK: The Boydell Press.
- Marwick, J.D.** (1894). *Charters and Documents relating to the City of Glasgow 1175-1649*. In British History Online. Online at <http://www.british-history.ac.uk/report.aspx?compid=47934> (Accessed 30 October, 2013).
- Mom** (2013). *Mom: Europe's virtual documents online*. Online at <http://monasterium.net/pages/en/home.php> (Accessed 30 October, 2013).
- PASE** (2010). *PASE: Prosopography of Anglo-Saxon England*. Online at <http://www.pase.ac.uk/index.html> (Accessed 30 October, 2013).
- Pasin, Michele and John Bradley** (2013). "Factoid-based prosopography and computer ontologies: Towards an integrated approach". In Literary and Linguistic Computing. published online June 29, 2013 doi:10.1093/linc/fqt037, and soon to be in print.
- PoMS** (2010). *Paradox of Medieval Scotland*. Online at <http://paradox.poms.ac.uk/> (Accessed 30 October, 2013).
- PoMS** (2013). *People of Medieval Scotland*. Online at <http://www.poms.ac.uk/> (Accessed 30 October, 2013).
- Vogeler, Georg** (2005). "Towards a Standard of Encoding Medieval Charters with XML". In Literary and Linguistic Computing. Vol 20 No 3. pp 269-280

An XML annotation schema for speech, thought and writing representation

Brunner, Annelen

annelen_brunner@gmx.de
Institut für deutsche Sprache

This contribution presents an XML schema for annotating a high level narratological category: speech, thought and writing representation (ST&WR). It focusses on two aspects: Firstly, the original schema is presented as an example for the challenge to encode a narrative feature in a structured and flexible way and secondly, ways of adapting this schema to TEI are considered, in order to make it usable for other, TEI-based projects.

The phenomenon ST&WR

ST&WR refers to the way the voice of a character is embedded in the narrator's text and is a feature that is present in most works of fiction. It has been widely studied in narratology, as it contributes to the construction of a fictional character, the narrator-character relationship and fictional world-building in general. Though ST&WR is partly defined by formal features like punctuation, verb mode, and sentence structure, narrative function is what is of interest in literary studies (cf.¹ for an overview). The challenge is to develop an annotation schema which is sufficiently structured to allow consistent annotation (especially with multiple annotators) and still captures nuances that are relevant for literary scholars.

The schema presented here – called ST&WR schema (ST&WR-S) – ties into literature studies as it uses categories agreed upon by most scholars and is similar to categorial systems proposed by narratologists Genette and Leech/Short (cf.^{2, 3}). The main influence was a project of Semino and Short, who annotated a corpus of English fictional, newspaper and

(auto)biographical texts for ST&WR with an SGML-conformant schema (cf.⁴).

ST&WR schema

ST&WR-S was developed for manual annotation of a corpus of 13 German narrative texts written between 1786 and 1917 (about 57 000 tokens). This corpus was then used as a reference for the development and evaluation of automatic methods for ST&WR recognition (cf.⁵). The purpose of ST&WR-S was twofold: It allows for a very fine-grained classification of ST&WR instances which is helpful in order to study the phenomenon and to do statistical studies on manually annotated data, like in Semino/Short's project. On the other hand it was designed to be modular and easily simplified to accommodate for the rougher classifications of automatic recognizers. Experiences during corpus annotation strongly influenced the design of the annotation schema.

ST&WR-S has three levels of specificity: Main categories, attributes and in some cases different values for further specifications of certain attributes. These are modelled as XML tags with attributes and values.

The manual annotation was done in the GATE framework for natural language processing (cf.⁶, <http://gate.ac.uk>). ST&WR-S is specified in XML schema files used by the plugin Schema_Annotation_Editor. Primarily, it is designed for inline XML, but GATE internally manages annotations as nodes and can convert them to a standoff XML format.

The main categories can be described with two axes: One axis represents the medium – speech, thought or written text (e.g. a quote from a character's letter). The second axis represents the four most common techniques of ST&WR: direct representation ("He said 'I am hungry.'"), free indirect representation ("Well, where would he get something to eat now?"), indirect representation ("He said that he was hungry."), and reported representation, which can be a mere mentioning of a speech, thought or writing act ("They talked about lunch."). This results in twelve main categories which are modelled as XML tags (*direct_speech*, *direct_thought*, etc.).

However, such a set of categories is necessarily rigid. When annotating a narrative phenomenon in a real corpus you will find many instances which are not clear-cut realisations of a predefined category. To deal with this fact, rather than just adding a confidence marker to the annotation, attributes are used to classify the type of deviation, so that the cases may be further studied and contrasted. As all attributes are optional and can be added to any main category, ST&WR-S allows for different levels of detail very easily. It is also possible to filter your annotation results afterwards by ignoring instances that carry a certain attribute.

Structurally, there is one numerical attribute (*level*), three attributes which are binary and just indicating whether the feature is present or not (*narr*, *prag*, *metaph*), two with optional further specification (*border*, *non-fact*) and one with mandatory further specification (*ambig*). All lists of attribute values are closed sets. Table 1 gives an overview.

Attribute name	Description	Values
<i>level</i>	level of embedment	numeric (default: 1)
<i>ambig</i>	ambiguity of the main category	Name of an alternative main category
<i>non-fact</i>	non-factual (eg. negated or hypothetical ST&WR) ("He did not admit that he loved her.")	<i>neg</i> , <i>hyp</i> , <i>fut</i> , <i>ques</i> , <i>imp</i> , <i>plan</i> , <i>unspec</i> (default: <i>unspec</i>)
<i>border</i>	borderline case of ST&WR ("He knew that he had lost.")	<i>percept</i> , <i>feel</i> , <i>state</i> , <i>unspec</i> (default: <i>unspec</i>)
<i>narr</i>	Ambiguity between ST&WR and non-verbal action ("She greeted her friends.")	binary (dummy value: yes)
<i>prag</i>	ST&WR, but with non-representational intent (e.g. politeness ("I suggest you leave now."))	binary (dummy value: yes)
<i>metaph</i>	metaphorical use ("His conscience told him to go.")	binary (dummy value: yes)

Functionally, *level* stands alone in the group as it does not mark a non-prototypical instance but is rather a 'monitor attribute'. It captures the level of embedment of a ST&WR instance, e.g. an instance of indirect thought that appears as part of an instance of direct speech would be tagged as *level=2*. This marker can then be used to study the behaviour of such embedded instances and compare their behaviour to non-embedded ones.

All other attributes deal with instances that deviate from the prototypical idea of ST&WR in relation to the definition of the main categories.

Ambig and *narr* both mark ambiguity. While *ambig* indicates that there is uncertainty as to which main category should be applied, *narr* signals that it is uncertain whether the instance is a case of ST&WR at all.

Border deals with uncertainty in regard to what is considered speech, thought and writing respectively. Especially thought representation is extremely tricky, as you have to decide what constitutes a thought. For example, the sentence "He knew he had lost." would be marked as *< indirect_thought border="state">*, as "to know" expresses a state of knowledge rather than a clear-cut thought. *Border* can also be applied to speech representation, e.g. if it is unclear whether there is a true verbalization like in the sentence "He screamed bloody murder."

Non-fact deals with instances where the ST&WR is non-factual and thus not a real 'representation' in the story world. Similarly, *prag* marks instances where ST&WR forms are used for non-representational purposes, especially politeness, and *metaph* represents metaphorical use of ST&WR.

In addition to that, the ST&WR-S contains two special categories modelled as XML tags. One is *frame*, which marks the framing clause of a direct representation which is not part of the representation itself but still interesting in the context of ST&WR. The other is called *embedded*. It can be used to mark embedded narratives which appear in direct representation (usually direct speech), e.g. if a character tells a story. Marking such cases with *embedded* essentially shifts the whole annotation level into a new narrative frame and gives it a different status than *direct_speech*. The use of *embedded* is optional and the tag can be easily transformed to *direct_speech* if this effect is not desired.

ST&WR-S is a valid XML schema but not compliant to TEI Guidelines. For sustainability it would be desirable to adapt it, as this would allow its usage in TEI-conformant documents without compromising their validity.

However, such an adaptation is not straightforward. The logical starting point is *<said>*, a tag from the quotation context which is defined for passages thought or spoken by real people or fictional characters (cf. ⁷). Though *<said>* is clearly intended to capture instances of ST&WR, its scope is narrower than the instances covered by ST&WR-S. In its core form, it only carries the attributes *aloud* and *direct*, both specified by truth values. *Aloud* is designed to distinguish between silent thought and passages spoken aloud (speech), but does not accommodate writing representation. *Direct* does not allow for any distinction between the ST&WR categories free indirect, indirect and reported. Of course, the rich attribute system of ST&WR-S does not have a predefined equivalent in TEI, either.

Several possibilities are considered how to adapt ST&WR-S while conserving its power as well as its modularity as much as possible. Ideas include use of standoff markup, possibly via the ** tag, modelling of the complex categorizations via feature structures, referenced by the *@ana* attribute, or extensibility of existing TEI-tags (most likely *<said>*).

References

1. McHale, B (2013). *Speech Representation*, in: Hühn, Peter et al. (eds.): The living handbook of narratology, Hamburg: Hamburg University Press, 2013 URL: www.lhn.uni-hamburg.de/article/speech-representation (last checked 17.10.2013).
2. Genette, G. (1980). *Narrative discourse. An Essay in Method*, Oxford: Blackwell.
3. Leech, G. and Short, M. (2007). *Style in fiction. A Linguistic Introduction to English Fictional Prose* 2. ed., London: Pearson Education Limited.
4. Semino, E. and Short, M. (2004). *Corpus stylistics*. Speech, writing and thought presentation in a corpus of English writing, London/New York: Routledge.
5. Brunner, A (2013). *Automatic recognition of speech, thought, and writing representation in German narrative texts*. Literary and Linguistic Computing 2013; doi: 10.1093/lil/fqt024
6. Cunningham, H. et al. (2011). *Text Processing with GATE* (Version 6): www.tinyurl/gatebook (last checked 01.11.2013).
7. Text Encoding Initiative (2013). *P5: Guidelines for Electronic Text Encoding and Interchange*, URL: www.tei-c.org/Guidelines/P5 (last checked 31.10.2013)

Europe as a Digital Network: EGO European History Online

Burch, Thomas

burch@uni-trier.de

Trier Center for Digital Humanities, Germany

Berger, Joachim

berger@ieg-mainz.de

Leibniz Institute of European History (IEG), Germany

The proposed paper reports on the development of a new digital resource: EGO | European History Online (EGO) is a **transcultural history of early modern and modern Europe** concentrating on processes of communication, interaction and interdependency (www.ieg-ego.eu). It is being published by the Leibniz Institute of European History (IEG) in Mainz in cooperation with the University of Trier's Centre for Digital Humanities. At its heart are transfer processes that extended across individual, familial and local realms and had a long-term impact. EGO traces these transfer processes in and between, amongst others, the realms of religion, law, politics, art, music, literature, economics, technology and the military, science and medicine. One can speak of a "transfer" when people, objects

and ideas move between different cultures (interpretative systems) and as a result undergo transformation.

European History Online is the first history of Europe that links the medium's relevance to the subject matter. The format of an online system of publication is the ideal medium for representing the complexity and dynamics of European communicative and transfer processes. The more than 200 articles are organised into ten thematic threads. These threads group the separate articles into a modular structure arranged thematically and methodologically. These threads are **transdisciplinary and multi-thematic**; they bring together the perspectives of different historical disciplines and their international authors. At the same time, they are organised diachronically, i.e. they deal with phenomena that – with specific periods of development and significance – are primarily evident throughout modern European history.

This organisation offers flexible means of accessing the contributions: in contrast to a printed book, *European History Online* does not have a beginning and an end. EGO accommodates the dynamics of intensifying communication and the continuously shifting intersections in European history by assigning many articles to more than one section. Within these multiple classifications, one can see how the topics interconnect. The different forms of presentation – surveys, basic elements and focus elements – and their organisation into a modular structure enable **nuanced contextualisation**.

In addition to this multi-layered structure, EGO articles are directly connected via hyperlinks. The aim of these connections is to expose the so far unknown concentrations of communication in European history, inspire new transcultural research in the various disciplines and thereby promote a more **dynamic understanding of European history**. The versatile search function allows users to put together their own "history of Europe" which corresponds to their individual interests.

Moreover, EGO pursues a **multilingual approach** that acknowledges the need for a workable meta-language / lingua franca in the Humanities but at the same time does justice to the linguistic variety of national academic cultures in Europe: EGO-articles are accepted in English and German. All major contributions are translated by native speakers and published in both languages. In addition, authors may publish their article in their native language. Users are invited to consult both the original and the translation in order to trace differing argumentative patterns and conceptual peculiarities of the respective languages.

From the technical point of view EGO is based on a sophisticated infrastructure which on the one hand supports the editorial board (backend of the system), and on the other hand is used to build the frontend of the system for the publication of the EGO articles in the internet. For these purposes the open source enterprise content management system Plone is used, which is a very powerful environment developed for professional use in organisations and companies. Especially its sophisticated and safe user access management as well as its workflow driven content management makes it one of the outstanding web management systems.

In Plone it is possible to handle very different content types as for example texts, images, PDF-files, audio and video data. For EGO all these project specific media types are modelled by a corresponding Plone article type, such that they can be configured by the editor himself within the Plone configuration layer. Here the editors can specify all relevant visual and layout attributes for each object according to the underlying web design of EGO and thereby prepare the article for the final publication. This step is integrated in the whole publication workflow which is modelled within the content management system. The participating roles are the authors, the editors, the copy editors and the publishers. Every member of this team has its own role and permissions to work with the documents, which run through different stages from 'private' (i.e. newly provided by the author) over 'review/revision' (redacted by at least one editor and/or copy editor) and 'internal preview' until it is 'externally published' on the EGO platform (c.f. figure 1).

The screenshot shows the EGO homepage with a sidebar for 'Theories and Methods' including 'Comparative History', 'Cultural Transfer', 'Transnational History', 'Transcultural History', 'Postcolonial Studies', 'History of Ideas', 'Knowledge Transfer', 'Historical Meso-Region', 'Mental Maps', and 'Europe'. The main content area displays the article's title, author (Fritjof Benjamin Schenk), original in German, displayed in English, and publication date (2013-07-08). It includes a table of contents, several small thumbnail images related to the article, and a sidebar with links to 'Historical Meso-Region', 'Europe', 'Balkan', 'Tourism', 'Edward C. Tolman (1886–1959) VIAP', 'Ottos B.', 'Cognitive Maps in Rats and Men', 'The West', 'From the "Turkish Mosaic" to Orientalism', and 'Johann Reinhold (1726–1786) und Georg Forster (1754–1794) im Denk'. A note at the bottom discusses the concept of 'spatial turn' in the humanities.

Fig. 1: EGO article 'Mental Maps: The Cognitive Mapping of the Continent as an Object of research of European History'

All information about the documents, the media types, the interlinking between documents and external resources and especially about the people involved is managed in an object database, which serves as storage for the backend as well as for the frontend. This guarantees that all changes to a document are logged in one consistent pool of data. Moreover, on this database the search engine for EGO is installed. In addition to a conventional full text retrieval the engine supports the query for specific categories of EGO articles as for example authors, time ranges, themes/threads, geographical regions of Europe and for media types (image, audio, video etc.).

To address aspects of long term availability of the project results the EGO documents are encoded according to international standards (XML/TEI, METS/MODS). The implementation of an open access interface to our Plone system allows a standardized export of the data, which is on the one hand used for long term storage in form of a dark archive. On the other hand it provides possibilities for the reuse of the EGO documents (e.g. linguistic analysis, visualization of the semantic network, compilation of bibliographic information, etc.).

The consequent use of a technical research and publication infrastructure leads to a highly dynamical publication process, to a very dense network of information and thereby to a very flexible presentation with additional benefits compared to traditional printed editions.

This change in media undertaken by *European History Online* challenges the concept of multi-volume published surveys, which, as a rule, must wait 20 to 50 years for a new edition. This **dynamic form of publication** corresponds to the dynamic understanding of Europe: the articles can be updated regularly, and the system can be extended by new articles in order to keep up with new developments in the research. Older versions of an article will remain accessible.

European History Online primarily uses linear, textual presentations of narrative and analysis in order to portray transfer processes in European history. However, EGO enhances **spatial perspectives** as well. All place names in the articles are georeferenced. They are retrievable in an alphabetical index and being visualized on a dynamic map (via Open Street Map). Thus authors are encouraged reflect their coverage of geographic areas in their research, whereas users, via the general index (and map) of place names, may discern spatial clusters with regard to their topic of interest. EGO thus strives to enhances a spatially-oriented approach in transnational and transcultural history.

In addition, *European History Online* combines **different types of media** in a – new – interpretative context. Images and audio and visual clips illustrate not only the topic being described, but also narrate their own histories of transfer and

enable new interconnections. EGO's transdisciplinary approach is also to a large degree a product of the images, graphics, maps, tables, film clips and audio samples linked to the different textual contributions. This network exists, on the one hand, via internal links to elements published within EGO and, on the other, via links to external images, textual sources and biographical data digitalised or published elsewhere, as well as – in the notes – scholarly literature and other academic resources online. While these external resources represent all national traditions relevant to the history of Europe, EGO makes them accessible to a transnational academic community via a **bilingual user interface**. The dynamic EGO system thereby brings together and groups thematically the range of international online resources on European history.

Socially-Derived Linking and Data Sharing within a Virtual Laboratory for the Humanities

Burrows, Toby

toby.burrows@uwa.edu.au
University of Western Australia

Verhoeven, Deb

deb.verhoeven@deakin.edu.au
Deakin University

Hawker, Alex

alex.hawker@versi.edu.au
VeRSI

HuNI (Humanities Networked Infrastructure) is a major new digital service for humanities researchers. Developed in Australia, with funding from the NeCTAR (National eResearch Collaborative Tools and Resources) programme, it aggregates data from 28 different cultural datasets from a variety of disciplines, makes them available for external re-use through an API and as Linked Open Data, and provides a set of tools for researchers to work with the data. HuNI is a virtual laboratory application which will be of great value to anyone interested in understanding Australia's rich cultural heritage. It is also the single largest aggregation of networked humanities data in Australia – a new national data service which is of cultural significance in its own right, and accessible to all. HuNI is planned to go into production in the second quarter of 2014.

We reported on the initial design stage of this project at a previous Digital Humanities conference.¹ This new paper will tell the full story of HuNI and will analyse, describe and evaluate its first full public release. It will include a demonstration of the full functionality of the service, and will report on its uptake by researchers and the wider community. We will also discuss the lessons learned from this large-scale project over its two-year lifespan, and the measures taken to ensure the sustainability of the HuNI service beyond the life of the project.

The paper will focus, in particular, on the ways in which HuNI is changing the nature of humanities research in the areas of data sharing, collaboration, community involvement, and the creation of socially linked data. Socially-derived linking of data is one of the key features of HuNI. Researchers are able to make assertions about relationships between entities represented in the aggregated data. If, for example, they search the data aggregate and identify two entities in their result set which are related in some way, they can add a link between the two records and define the nature of the relationship. The linking statement may be drawn from an existing vocabulary of relationships, or may use free text entered by the user. The virtual laboratory also allows a researcher to assert that two entities are not related, in recognition that this kind of statement is also a key characteristic of humanities research.

To help visualize these social links, each entity has its own network graph, showing up to six degrees of separation, resulting in a growing network of dynamic connections, or

the “networked effect”. These “social linking” assertions are visible in the HuNI data aggregate. They may also appear in virtual collections assembled and published by individual users of the HuNI virtual laboratory. These virtual collections can be published for other users to see and re-use. Some links between entities have also been imported from the source datasets as part of the harvesting and ingest process. HuNI users can annotate these links with their own assertions as well.

Crucially, the provenance of all these “social linking” statements is also captured, enabling subsequent researchers to see who made each assertion. HuNI is an aggregate with a sense of its own history, in which researchers can trace how records have changed over time. Humanities research not only involves making connections between entities; it also involves assessing any changes in cultural flows and network relationships through time. So each HuNI record is timestamped, meaning that researchers will always see the current view of a record, with its related records and assertions, but will also have the option to view how the record has changed since it was first harvested. The provenance information for each record, together with any curated assertions, is captured, so that researchers can see when the records were harvested and by whom. A link to the originating data record at source is also provided in the user interface.

“Social linking” is a crucial feature of the HuNI virtual laboratory. Instead of relying on a pre-coordinated mapping to a detailed ontology, we are relying on researchers and community users to establish most of the connections within the heterogeneous data aggregate. This enables HuNI to capture the different disciplinary perspectives of users, rather than trying to fit all the data into a single normative framework. It also acknowledges the productive differences that both define and link specific domains through a form of generative knowledge transfer. The opportunity to link data socially encourages HuNI users to share their knowledge and research findings in the form of specific assertions, and to discuss these statements with each other. In the paper, we will report on the ways in which this feature is being used by researchers and community users, and the extent to which it is enabling a new approach to data sharing in the humanities.

We acknowledge the contributions of Dr Marco La Rosa (Solution Architect) and Dr Anne Cregan (Semantic Lead and Business Analyst) in developing the HuNI infrastructure.

References

Burrows, T. (2012). *Designing a national 'Virtual Laboratory' for the humanities: the Australian HuNI project*. In Meyer, J. C. (ed), Digital Humanities 2012: Conference Abstracts. held July 16-22 at University of Hamburg. 139-141

SyMoGIH project and Geo-Larhra: A method and a collaborative platform for a digital historical atlas

Butez, Claire-Charlotte

charlotte.butez@ish-lyon.cnrs.fr
LARHRA UMR 5190, France

Beretta, Francesco

francesco.beretta@ish-lyon.cnrs.fr
LARHRA UMR 5190, France

The aim of this paper is to highlight the need for digital atlases in historical research and to present the data model and the collaborative platform we have developed in order to produce a historical geographic information system (HGIS), the Geo-Larhra, which is suitable for producing a new digital historical atlas.

1. Background and purposes

At its beginning the project explored the possibilities of integrating geographical and historical data into the same digital research platform. Several reasons fuelled this project. The main issue that arose was how to obtain base maps to represent a specific historical time. Traditionally, printed historical maps and atlases provide spatialized maps usually connected with relevant historical dates. This chronological selection documents major historical events for a specific geographic area (e.g. the political borders of European Countries after the Congress of Vienna, 1815). More rarely, maps are provided by century (1600, 1700, etc.) and significant elements in the chronological development or specific changes in tight spaces do not appear. Moreover, there exists a limited range of digital historical maps, particularly with regard to freely accessible shapefiles (commonly used geospatial vector data format), and they often do not take into account the diachronic changes in political or administrative territories. The conclusion was clear, the advance of digital history needs a project for realizing a new digital historical atlas, enhanced by researchers working collaboratively in a coherent, easy-to-use environment.

Drawing on different successful experiences in historical atlases (e.g. A vision of Britain through time, Euratlas, Digitaler Atlas zur Geschichte Europas seit 1500, HGIS Germany, China Historical GIS), we have devised a method for modeling geohistorical data to process the evolution of territories. This work is an application of the SyMoGIH method, the MOdular SYstem for Historical Information Management (Beretta, Francesco / Vernus, Pierre (2012)). Against this background, the historical atlas will be available to historians participating in the SyMoGIH project but it will also be accessible to a larger public through the web site, www.geo-larhra.org, offering a basic mapping service online as well as downloadable resources. Geo-Larhra includes a gazetteer, a catalog of vector layers plus the digital historical atlas. In this paper we will describe the underlying data model and the principles and workflow of our collaborative approach.

2. Method and data model

To build such a collaborative historical atlas, we needed to develop a generic data model allowing the processing of any type of place and taking into account any kind of temporal evolution due to the toponymical, typological or spatial extent changes. To address the issue of multi-dimensional evolution, the team of the SyMoGIH project has developed a generic data model independent of any research problem [see the documentation on our website: www.symogih.org/?q=documentation]. In our model, we distinguish between the identification of places and their spatial representation.

The identification of places is carried out in the traditional way using a gazetteer: a place is identified by its name or names, a type and a geographical location in form of a point or a bounding box (cf. Hill, Linda L., 2006). Each place is identified by a uniform resource identifier (URI).

The processing of spatial representation on the contrary is the most novel part of our method : we have introduced a distinction between the form of a place at a given time, that we call a “concrete time-specific form” or simply a “concrete form”, and the more or less accurate geometries (i.e. geo-referenced vector data) representing this form at different scales. The evolution of the place's form is first described and documented by historical information collected collaboratively by the historians participating in the project. The geometries are then produced by the GIS specialists according to the collected information. This modeling process and data production workflow is more flexible and suitable for historical research than the traditional method in GIS, which links data directly to geometries. By using SQL and spatial queries it is possible to output the shape of places and territories at a given date with a temporal scale which is currently accurate to the day. This method leads to a synthesis between combining the traditional practice of historical databases, the use of historical atlases and GIS methods. Geo-Larhra is intended to be a resource to address both of these needs : providing historically accurate

base maps and allowing historians to make spatial analysis taking into account temporal evolution.

Our philosophy is based on a collaborative and open approach aimed at enriching and developing the historical atlas. The collaboration of historians and GIS specialists is carried out on several levels : they collect historical and geographical information from sources, maps and historical atlases ; they produce historical data in the collaborative database platform ; they produce geometries using the collected historical and geographical data. Digital maps are finally created and they can be successively added to, following the same workflow, if new or more precise information is collected. The paper will give some concrete examples of this collaborative approach.



Fig. 1: Screenshot of the historical atlas : www.geo-larhra.org

3. The platform

Technological choices for the project encountered strong constraints resulting from the collaborative aspect and the generic system and multiple uses for which it was intended. The software architecture has been constructed using the triptych PostgreSQL, Post-GIS and QGIS. The DBMS PostgreSQL provides several advantages. It is a free and open source, and useful to establish the precise management of users' rights. The PostGIS extension is easily interfaced with other management database tools to query, analyze and visualize data (GIS software, statistical analysis, GIS web server). Geo-historical data are published with the TinyOWS map-server which provide WMS (Web Map Service) and WFS-T (Web Feature Service) in QGIS or OpenLayers.

4. Future prospects

To date, the data model presented in this paper seems to fit historians' needs perfectly. However our team must now improve the ease of use of the platform accessed by scholars concerned by spatial analysis who would contribute to this project. We have already started to publish the gazetteer on the web and we provide some shapefiles, extracted from the historical atlas of the Italian peninsula territories, which was our first data set created according to this method. Our longer-term perspective is to expand the geographical area of the atlas with the help of international partners.

References

- Beretta, F., Vernus, P.** (2012). "Le projet SyMoGIH et la modélisation de l'information : une opération scientifique au service de l'histoire", Les carnets du LARHRA, 1, 81-107
- Beretta, F., Butez, C.** (2013). Conception et usage du système d'information géo-historique de SyMoGIH. Partie 2 : Exemple d'application : conception de l'atlas historique numérique et analyses de données attributaires de l'Italie du Risorgimento Géomatique Expert, n° 92, Mai-Juin 2013, pp. 48-54.
- Butez, C.**, (2013). Conception et usage du système d'information géo-historique de SyMoGIH. Partie 1: Naissance et conception d'un système d'information géo-historique collaboratif Géomatique Expert, n° 91, Mars-Avril 2013, pp. 30-35.
- Elli, P.S., Gregory, I. N.** (2007). *Historical GIS : technologies, methodologies, and scholarship*, Cambridge/New York , Cambridge University Press
- Gregory, I. N.** (2003). *A place in history : a guide to using GIS in historical research*, Oxford, Oxbow Books
- Hill, L. L.** (2006). *Georeferencing : the geographic associations of information*, Cambridge Mass, The MIT Press

Towards visualizing linguistic patterns of deliberation: a case study of the S21 arbitration

Bögel, Tina

tina.boegel@uni-konstanz.de
Universität Konstanz

Gold, Valentin

valentin.gold@uni-konstanz.de
Universität Konstanz

Hautli-Janisz, Annette

annette.hautli@uni-konstanz.de
Universität Konstanz

Rohrdantz, Christian

christian.rohrdantz@uni-konstanz.de
Universität Konstanz

Sulger, Sebastian

sebastian.sulger@uni-konstanz.de
Universität Konstanz

Butt, Miriam

miriam.butt@uni-konstanz.de
Universität Konstanz

Holzinger, Katharina

katharina.holzinger@uni-konstanz.de
Universität Konstanz

Keim, Daniel A.

daniel.keim@uni-konstanz.de
Universität Konstanz

1. Introduction

This paper reports on the interdisciplinary project VisArgue which is concerned with the automatic linguistic and visual analysis of political discourse with a particular focus on the concept of deliberative communication^{1 2 3 4 5}. According to the theory of deliberative argumentation, stakeholders participating in a multilog, i.e. a multi-party conversation, should justify their positions truthfully and rationally and should eventually defer to the better argument.

Automatically measuring the deliberative quality of a multilog calls for an identification of linguistic cues that shed light on issues such as objective vs. subjective argumentation, invocation of the common good or democratic notions as part of the argument. Notions such as speaker stance, speaker belief/certainty are also immediately relevant, as is an analysis of rhetorical devices known to trigger conventional implicatures⁶. In short, a promising way of arriving at an operationalization of the indications for the deliberative quality of a multilog is a linguistic analysis of the linguistic cues present in the multilog.

This paper presents on-going work on analyzing strategies for argumentation via automatic means, using public data from a German arbitration. In addition to a thorough linguistic analysis of the relevant parameters, we provide a computational implementation that automatically annotates the corpus with respect to pragmatically relevant features. This implementation combines a rule-based system that reflects deep linguistic

analysis with a visual analysis system that also provides results with respect to more shallow natural language processing methods such as keyword identification, topic modeling, and standard calculations with respect to length of utterances, amount and type of turn-taking, etc.

2. Data

In Germany, the method of deliberative discourse has been increasingly applied to the resolution of large-scale public conflicts since the early 1990s. One recent well-known example is the public arbitration process on "Stuttgart 21" (S21), a new railway and urban development project in the city of Stuttgart. In response to massive public criticism of the project, a public arbitration procedure was established. The data for our initial investigation are the transcribed minutes of the S21 arbitration process, which consist of nine days of sessions, each lasting for around seven hours, with a total of about 70 different speakers. The transcripts consist of spoken German conversation between mediator, experts, project supporters and opponents and are converted into an XML-readable format in order to facilitate later processing and annotation. Based on the information contained in the web transcripts, the XML transcripts are annotated with speaker information and the general topic of the session. Overall, the transcripts contain around 265.000 tokens in 1330 utterances.

In order to arrive at a more fine-grained analysis of the discourse, all utterances have to be split up into elementary discourse units (EDUs)⁷. Although there is no consensus in the literature on what EDUs are, in general, each DU is assumed to describe a single event (e.g.,⁸). In the case at hand, we approximate this by treating all lexical items between two punctuation marks as belonging to one DU, a method that is commonly assumed in discourse parsing.

3. Linguistic background

A central aspect of our work is a linguistically motivated operationalization of features that indicate the deliberative quality of a multilog, important parameters being the realization and the communicative function of arguments in the discourse as well as speaker stance and speaker belief. We have decided to initially focus on just two of the relevant linguistic parameters found in German: the interaction between causal discourse connectors and modal particles.

Causal discourse connectors (e.g., *da*, *weil*, *denn*, *zumal* 'because (of)/due to/as') generally introduce a justification of a speaker's statement (e.g.,⁹). These connectives and the justification they indicate can be extracted automatically. However, the precise shade and force of the argument being made, including speaker stance and speaker belief are modulated in spoken German by a heavy use of modal particles (e.g.,^{10 11}). For instance *halt* or *eben* indicate a conventional implicature that the speaker believes the argument to refer to an immutable constraint imposed by the outside world, exemplified in (1). The particle *ja* in (2), in contrast, signals that the speaker assumes that the content of the argument is part of the common ground of the multilog participants.

(1) [...] weil halt in dem Bereich die meisten Autos unterwegs sind.

[...] as HALT in Art area Art most car.PI underway be.3.PI

'[...] because most cars are underway in this area.' (Dr. Heiner Geissler, S21, Nov. 4th 2010)

(2) [...] da Sie ja gesagt haben, dass [...] as Pron.2.Sg.Pol JA say.Past.Part have.Inf that [...] as you JA said that [...] (Tanja Gönner, S21, Nov. 4th 2010)

3.1. Ambiguity

Ambiguity presents a serious problem for the automatic extraction and identification of both causal connectors and modal particles. For example, especially *da* 'as' presents a challenge for automatic processing, because of its multi-functional usage as either the temporal or locative pronoun 'there' or as a connector meaning 'because'. However, such ambiguities can be largely resolved by taking linguistic factors such as the position of the connector, its neighboring elements and the general structure of the carrier sentence into account. In (3), we schematize the identification rule for the German causal connector *da* 'as'.

(3) IF *da* not followed by verb AND
da not preceded by a particle or another causal connector AND
final verb is an infinitive THEN
da is a causal connector.

The same procedure is followed with respect to modal particles such as *eben*, which can be a focus particle, a temporal adverbial meaning 'just', or a modal particle that indicates the speaker's resigned acceptance of a fact due to an immutable constraint¹².

3.2. Inference rules

While these two dimensions are by themselves important for the interpretation of a given discourse, the additional benefit for measuring deliberation results from a combination of the two dimensions. Taking the example in (1), the inference rule in (4) yields the annotation in Figure 1.

(4) IF causal connector found AND causal connector followed by particle denoting immutable constraint THEN
annotate the DU start tag with <DiscRel="justification" CI="immutable_constraint">

```
<discourse_unit id="17" DiscRel="justification" CI="immutable_constraint">
<lexeme id="1" connector="causal">weil</lexeme>
<lexeme id="2" particle="resignation_acceptance">halt</lexeme>
<lexeme id="3">in</lexeme>
<lexeme id="4">dem</lexeme>
<lexeme id="5">Bereich</lexeme>
<lexeme id="6">auch</lexeme>
<lexeme id="7">die</lexeme>
<lexeme id="8">meisten</lexeme>
<lexeme id="9">Autos</lexeme>
<lexeme id="10">unterwegs</lexeme>
<lexeme id="11">sind</lexeme>
</discourse_unit>
```

Fig. 1: Annotation of example (1).

On the other hand, the rule in (5) deals with the combination of the causal connector *da* and the modal particle *ja*, rendering the annotation of example (2) in Figure 2. (5) IF *da* is used as causal connector AND *da* is followed by particle denoting common ground THEN
annotate the DU start tag with <DiscRel="justification" CI="common_ground">.

```
<discourse_unit id="2" DiscRel="justification" CI="common_ground">
<lexeme id="1" connector="causal">da</lexeme>
<lexeme id="2">Sie</lexeme>
<lexeme id="3" particle="common_ground">ja</lexeme>
<lexeme id="4">gesagt</lexeme>
<lexeme id="5">haben</lexeme>
</discourse_unit>
```

Fig. 2: Annotation of example (2).

These inferences, which perform context-sensitive linguistic annotation of discourse units, help to interpret the whole discourse and shed light on the way speakers and listeners interact, incorporating detailed linguistic knowledge about syntax and discourse pragmatics.

Despite the comparatively small corpus, it is nevertheless difficult to see overall patterns of argumentativity at a glance, while still maintaining a detailed view on single annotations. In order to overcome this drawback, we introduce a visualization system which encodes those annotations visually and makes the patterns more accessible.

4. Visualizing argumentativity

The visualization of linguistic patterns has been shown to shed light on a number of phenomena, from theoretically motivated topics like phonological patterns¹³ and lexical semantic change¹⁴¹⁵ to machine learning issues with respect to clustering¹⁶. The goal of visualizing the structure of argumentativity across the discourse is twofold: First, patterns of argumentation that have been identified through the linguistic inference rules can be analyzed in their context. Second, the distribution of arguments over the course of the conversation may reveal additional knowledge on the deliberative quality of different parts of the overall discourse.

Figure 3 shows a visualization of parts of the S21 arbitration session on Nov. 4th 2010, each sentence occupying one line, each speaker turn contained in a grey square. The bars marked in yellow represent discourse units containing justifying statements. The tool is interactive in that the user can zoom in and out of the discourse and can investigate the relevant discourse units in detail without loosing the overall distribution. A detailed view on the data is shown in Figure 4.

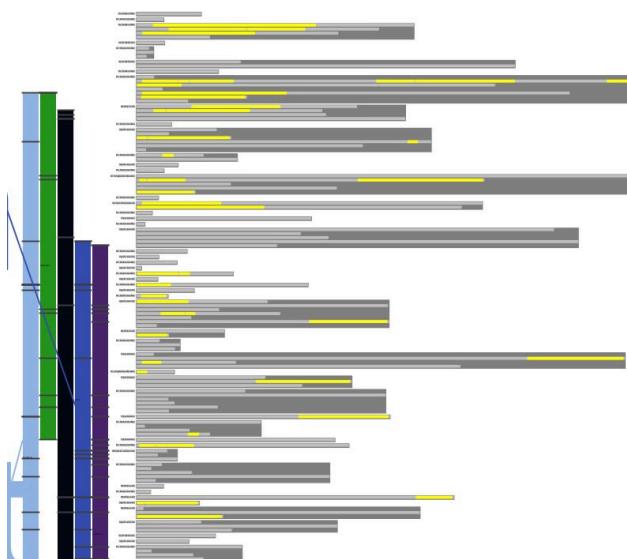


Fig. 3: Visualization of justifying statements in the S21 arbitration on November 4th, 2010.

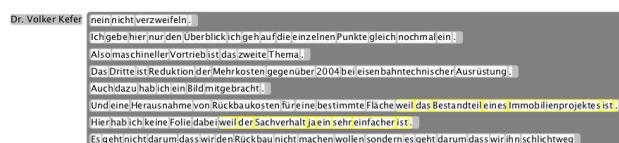


Fig. 4: Detailed visualization of justificational discourse units.

5. Summary and future work

This paper presents an approach of operationalizing the notion of deliberation using discourse connectors and modal particles in order to shed light on the way arguments are exchanged and how speakers and listeners relate to them. By using a visualization approach, the annotated data can be inspected over the whole discourse, allowing for an interpretation of the role that argumentativity plays in the arbitration. In the future, we will incorporate more linguistic cues that are relevant for deliberative communication and also deal with multiword instances that are relevant on a number of levels. With the increasing number of annotation levels, the visualization will be extended to show the interactions between different levels, allowing for more insights into discourse structure and eventually deliberation.

Acknowledgements

We thank Mennatallah el Assady, Manuel Hotz and Rita Sevastjanova for their help with implementing the discourse visualization and the German Ministry for Education and Research (BMBF) for their funding under the eHumanities research grant 01UG1246.

References

1. Habermas, Jürgen. (1981). *Theorie des kommunikativen Handelns*. 2 Bände. Frankfurt am Main: Suhrkamp.
2. Habermas, Jürgen. (1992). *Faktizität und Geltung: Beiträge zur Diskurstheorie des Rechts und des demokratischen Rechtsstaates*. Frankfurt am Main: Suhrkamp.
3. Dryzek, John S. (1990). *Discursive Democracy: Politics, Policy and Political Science*. Cambridge: Cambridge University Press.
4. Bohman, James. (1996). *Public Deliberation: Pluralism, Complexity and Democracy*. Cambridge, MA: The MIT Press.
5. Gutmann, Amy and Dennis Thompson. (1996). *Democracy and Disagreement — Why moral conflict cannot be avoided in politics, and what should be done about it*. Cambridge, MA: Harvard University Press.
6. Potts, Christopher. (2012). *Conventional implicature and expressive content*. In Claudia Maienborn, Klaus von Heusinger, and Paul Portner, eds., *Semantics: An International Handbook of Natural Language Meaning*, Volume 3, 2516-2536. Berlin: Mouton de Gruyter.
7. Marcu, Daniel. (2000). *The Theory and Practice of Discourse Parsing and Summarization*. Cambridge, MA: The MIT Press.
8. Polanyi, Livia, Martin van den Berg, Chris Culy, Gian Lorenzo Thione, David Ahn. (2004). *Sentential Structure and Discourse Parsing*. Proceedings of the ACL2004 Workshop on Discourse Annotation.
9. Prasad, Rashmi, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi & Bonnie Webber. (2008). *The Penn Discourse Treebank 2.0*. Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008). Marrakech, Morocco.
10. Zimmermann, Malte. (2011). *Discourse Particles*. In P. Portner, C. Maienborn und K. von Heusinger (eds.), *Semantics. (Handbücher zur Sprach- und Kommunikationswissenschaft HSK 33.2)*. Berlin, Mouton de Gruyter. 2011-2038.
11. Karagjosa, Elena. (2004). *The Meaning and Function of German Modal Particles*. Saarbrücken Dissertations in Computational Linguistics and Language Technology.
12. Min-Jae Kwon (2005). *Modalpartikeln und Satzmodus*. Untersuchungen zur Syntax, Semantik und Pragmatik der deutschen Modalpartikeln. Dissertation, LMU Muenchen.
13. Mayer, Thomas, Christian Rohrdantz. (2013). *PhonMatrix: Visualizing co-occurrence constraints of sounds*. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 73-78.
14. Rohrdantz, Christian, Annette Hautli, Thomas Mayer, Miriam Butt, Daniel A. Keim and Frans Plank. (2011). *Towards Tracking Semantic Change By Visual Analytics*. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pages 305-310.
15. Rohrdantz, Christian, Andreas Niekler, Annette Hautli, Miriam Butt and Daniel A. Keim. (2012). *Lexical Semantics and Distribution of Suffixes - A Visual Analysis*. In Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH, pages 7-15.
16. Lamprecht, Andreas, Annette Hautli, Christian Rohrdantz and Tina Bögel. (2013). *A Visual Analytics System for Cluster Exploration*. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 109-114.

Mining poetic rhythm: using text-to-speech software to rewrite English literary history

Cade-Stewart, Michael

michael.cade-stewart@kcl.ac.uk
Kings College London

1.1. Overview

Automatically predicting where the beats fall in a line of English poetry is difficult. This is because the stress placed on a word will be dependent upon the meaning of the statement of which it is a part. Without approximating the meaning of a line, then, it is not possible to arrive at an accurate prediction as to where the beats will fall when the poem is read or performed. This is no trivial task. Yet without determining the beats, it is impossible to identify the metre of the line or its sound-patterning, such as alliteration and assonance, which are instrumental to the rhythm. Because of this difficulty, the history of English poetic rhythm remains almost entirely uncharted. Twenty years ago, Preminger and Brogan's *New Princeton Encyclopaedia of Poetry and Poetics* lamented that 'there is at present no comprehensive and reliable history of the development of metrical practice in the West, not even competent accounts for any one language'.¹ The intervening decades have not resolved this problem for English. While there have been notable attempts to construct a reliable history, none have been comprehensive. Working without the benefit of computer automation, literary historians have focused exclusively on a small number of major poetic figures.

Having done so, the current histories exhibit erroneous assertions of innovations in rhythm and rhyme, give credit to the wrong poets, and neglect formal experiments occurring on the margins. Further, most approaches tend to rely on the impressions of the critic, rather than empirical data. Consequently, the majority of statements about poetic rhythm have been made in a position of ignorance about the formal characteristics of the corpus of English poetry.

The problem becomes acute from the start of the nineteenth century as the volume of printed material explodes. Nevertheless, Martin J. Duffell's *New History of English Metre* (2008) considers only 32 major poets for the period, analysing 200-300 lines of poetry as representative samples of their practice.² His analysis of poetic form therefore encompasses fewer than 9,000 lines. This is grossly inadequate, given that W. B. Yeats's *Collected Poems* alone surpasses this total by 2,000 lines. The number of poems considered by Duffell is a tiny proportion of the 151,299 poems in English considered important enough to be included in the Literature Online (LION) database for this period.

1.2. Methodology

The problem of coverage can be overcome by harnessing the power of 21st Century achievements in speech synthesis and text-to-speech software. My three-year project at King's College London, funded by the British Academy, is concerned with doing just this. It uses the MARY text-to-speech software, developed by the DFKI (the German Research Centre for Artificial Intelligence) and the Institute of Phonetics at Saarland University. The intended outcome of the project is a more reliable and comprehensive account of developments and trends in poetic rhythm, metrical forms, sound-patterning, rhyme schemes, and stanza types, in verse in English for the period of 1800-1970.

The MARY-tts software has recently been used to visualize sound patterns in literary texts by Tanya Clement et al, representing aural data as high-highlighted text.³ My approach differs by extracting key data from an intermediate stage of natural language processing that enable the software to identify some of the beats in a line of poetry. My scripts then extrapolate where the other beats are likely to fall, if the line be read as poetry rather than normal speech. From these data, my tools can distinguish between binary and ternary metres, determine the number of beats per line in the metrical template, and identify trochaic and iambic verse. It can spot refrains, and repeated structures, and sound-patterning such as internal

rhyme and alliteration. The approach has been tested on verse selected from a wide variety of sources, including Wordsworth, Browning, Keats, Tennyson, Hopkins, Eliot, and the entire corpus of Yeats's poems.

There have been earlier attempts to automatically determine formal qualities of poems using digital tools. Marc Plamondon's tool developed for the Representative Poetry Online database at Toronto remains an impressive example.⁴ However, none of these approaches have employed text-to-speech software to predict phonemic properties of the text automatically. The manual entry of syllables and phonemes for each word inevitably limits the scope and accuracy of such tools. Other impressive attempts to uncover the grammatical rules of particular poets' prosody have been similarly limited by manual input of language processing.⁵

1.3. Conclusion

My preliminary results suggest that the project will succeed in its aim of massively enlarging and enriching our understanding of literary history. Further, in assigning credit to the real progenitors of formal trends, and in providing the data with which to analyse formal developments empirically, rather than impressionistically, this project has the potential to rewrite the history of poetic form, and of poetic influence, altogether.

References

1. Preminger, Alex and Brogan, T. V. F and Warnke, Frank J. and Hardison, Jr., O. B. and Miner, Earl. *The New Princeton Encyclopedia of Poetry and Poetics*. Princeton University Press, Princeton, N. J.: 1993. p. 777.
2. Duffell, Martin J. (2008), *A New History of English Metre* Modern Humanities Research Association and Maney Publishing, London.
3. Plamondon, M. R. (2006). *Virtual Verse Analysis: Analysing Patterns in Poetry*. Literary and Linguistic Computing. 21. Supplemental Issue: 127-141.
4. Hayes, Bruce and Wilson, Colin and Shisko, Anne (2012). *Maxent grammars for the metrics of Shakespeare and Milton*. Language 88:4 (December 2012): 691-731.

The Landscapes of Casta Paintings: Depictions of Social Anxieties in XVIII Century New Spanish Art

Caldas, Natalia

ncaldas@uwo.ca
CulturePlex Lab, UWO

Ortega, Élika

eortegag@uwo.ca
CulturePlex Lab, UWO

Jiménez Mavillard, Antonio

ajimene6@uwo.ca
CulturePlex Lab, UWO

Brown, David

dbrow52@uwo.ca
CulturePlex Lab, UWO

Suárez, Juan Luis

jsuarez@uwo.ca
CulturePlex Lab, UWO

1. Introduction

Casta painting was one of the most popular non-religious artistic genres in New Spain (present day Mexico) during the XVIII century. They come in series of up to sixteen scenes,

each one showing an interracial couple and their offspring usually carrying out daily activities in everyday settings. In these paintings, artists depicted the three main ethnic groups making up Mexico's colonial population: *Españoles*, *Indios*, and *Negros*, and the process of *mestizaje*. Produced mostly in Mexico City and Puebla, Casta paintings reached a peak in production between 1770 and 1780, disappearing at the beginning of the 19th century as the War of Independence began and a generalized rejection of colonial structures took hold. In New Spanish society *casta* referred to race, both in biological and social terms¹. This notion was the basis of a caste system that pervaded New Spanish life at the time; a form of colonial control informing the kind of jobs people could do, where they could live, the civil liberties they had, and whether they paid taxes. According to Edward Long, the caste system had three general purposes: "first, to guarantee that each race occupy a social niche assigned by nature; second, to offer the possibility of improving one's blood through the right pattern of mixing; third, to inhibit the mixture of Indians and Blacks, which was deemed the more dangerous to the Spanish social order"². Blood mending—a process that was believed to make the offspring of interracial couples return to a pure European bloodline—was in the Spanish elite's eyes a way to assert their prominence. A control mechanism serving colonial concerns, blood mending was commonly staged in Casta paintings and, along with hierarchical and structured serialization, appears to suggest an ordered and stable social system³.

Critics have suggested that Casta paintings served as souvenirs—postcards—of the new world and, thus, showed a functioning and harmonious society⁴. As a matter of fact, Magali Carrera has stated that "as visualizations of race, Casta paintings stabilize the ambiguity and complexity of physical race by locating the meanings of race in the confluence, interactions, and mediations between and among physical, social, and economic spaces"⁵. Because of the composition and titles of Casta paintings, much literature has focused on their function as documents cataloguing the existence of the main races and the many resulting combinations—upwards of fifty according to Nicolás León⁶. We take the painting's titles as metadata rooted in the hegemonic perception of the groups depicted and explore their inconsistencies; for example, how in the *casta* system 'casta' simultaneously refers to both a specific mix like '*Barcino*' or '*Coyote*' and to the whole set of possible *mestizaje* instances. For Margarita de Orellana, the proliferation of terms and the representation of many *castas* showed the discontinuity of the *mestizaje* phenomenon. Furthermore, by dividing and fragmenting it, the complexity of *mestizaje* was over simplified⁷. The fact that many of Casta paintings were commissioned by and for Spanish patrons, and even meant to be sent to Europe⁸, explains the intent to reduce the complexity of *mestizaje* and to show the stability of life in the Americas. Moreover, the notion of blood mending in the caste system, observable in the paintings, sought to put an end to "the widely held notion in Europe that everybody in the Americas was hopelessly mixed"⁹. We argue that by artificially emphasizing the division of *castas*, these paintings sought to reassure the Spanish who feared their demise in New Spanish society, but ultimately failed to do so.

2. Overview and Methodology

In this study we challenge the hierarchical catalogue view embedded in Casta paintings. Instead of looking at *castas* in their rigid classifications, we have examined them as a group making up the larger and more complex figure of the *mestizo*. In that sense, we take *casta* as a group category referring to all *mestizos* resulting out of *Español*, *Indio*, and *Negro*. We base this assertion following Octavio Paz for whom, "among all the groups making up New Spain's populations, *mestizos* were the only ones embodying that society, its true children... Furthermore, they were those who made it not only new, but another" (our translation)¹⁰. In addition, instead of looking at particular series we have built a database with descriptions of over two hundred paintings. Painting description included extracting a total of 618 characters and marking them by *casta* (as assigned to them in the paintings' title), gender,

as "parent" or "child", and the activity they are doing. Casta painting critics have often pointed out that far from being a harmonious reflection of the caste system, they highlight the tensions and complexities underlying it. We build upon this through an extensive data approach and analysis. We believe this global approach has the capacity to show how, far from pacifying Spanish fears and concerns, Casta paintings confirm them and suggest prevalent *mestizaje* and the loss of Spanish prominence.

3. Results

Data analysis has shown that the majority of the characters depicted in Casta paintings are either *Españoles* or *Indios*, followed by *Mulatos* and *Mestizos* (Fig. 1). Most interesting is a composite view of *casta* and gender in which *Español* males are the most represented group throughout and are present in almost half of the paintings we have described. The second largest group is *Indio* females, present in over seventy paintings (Fig. 2). The recurrence of *Español* male and *Indio* females signals the basis of racial mixing resulting in the particular figure of the *Mestizo*, which can be extrapolated as the general view of *casta* as any of the mixes.

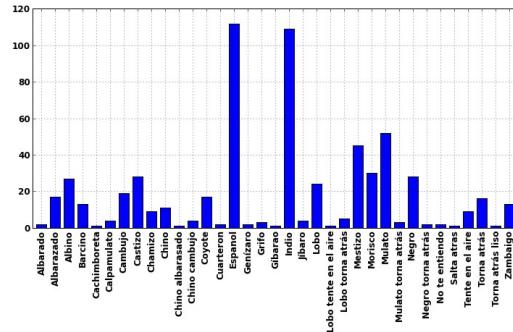


Fig. 1: Overall casta frequency.

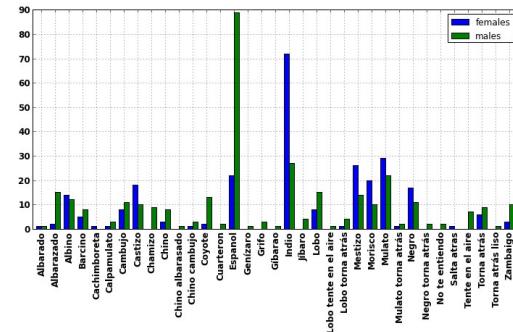


Fig. 2: Casta frequency by gender.

A look at the offspring sheds much light into the issues of blood mending and Spanish prominence. When we look at the offspring as separate *castas*, it is indeed the *Español* children who are most recurrent. The division of *castas*, thus, fulfills the objective of showing Spanish prominence among the interracial diversity. Furthermore, as there are no couples composed of both *Español* father and mother, the presence of *Español* offspring highlights the possibility of blood mending through the right combinations (Fig. 3). In contrast, when we look at offspring after *casta* grouping, the outlook is quite different. In Fig. 4, we show the proportion of two non-mixed offspring (*Español* and *Indio* as there are no *Negro* offspring in the corpus) and the two biggest *casta* groups (*Mestizo* and *Mulato*). This grouping highlights how, even though the *Españoles* are the larger minority, their prominence quickly dwindles in comparison to *casta* children. Fig. 5 reinforces the sweeping presence of mixed-race offspring in comparison to the *Español*. Our data on the offspring depicted suggest not only that blood mending and a return to whiteness are rather marginal, it also

highlights that further mixing is the standard depicted in Casta paintings.

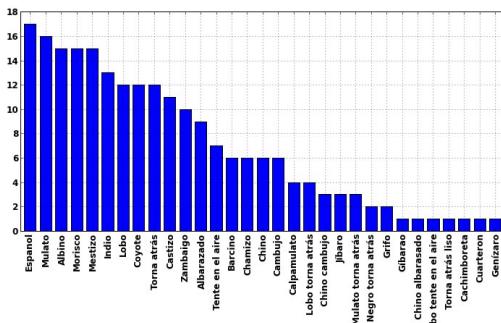


Fig. 3: Overall offspring casta frequency.

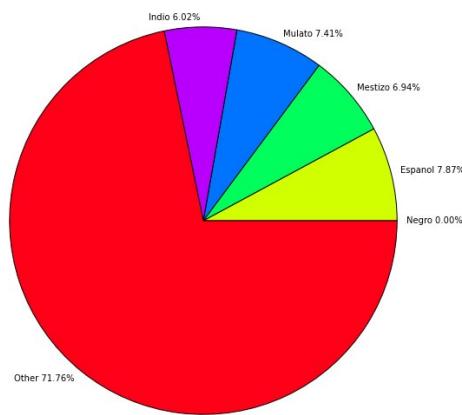


Fig. 4: Main castas and casta grouping.

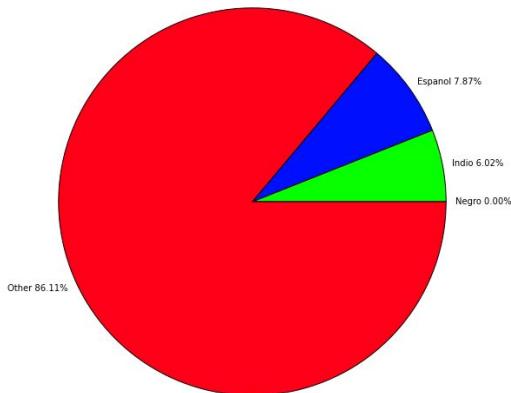


Fig. 5: Espanol, Indio, Negro and Castas outlook.

Additionally, we have identified forty activities being depicted in Casta paintings, and explored the three most recurrent ones in detail: carrying and holding children (Fig. 6), portrait posing (Fig. 7), and working (Figs. 8 and 9). We have looked at these

activities with regards to *casta* frequency. In terms of carrying and holding a child, females are seen to be the ones who carry out this activity the most; however, it is actually *Indio* women who have the highest frequency. For the males, it is seen that *Español* take charge and hold children over *Español* women, and over every other male. *Español* males are most frequently depicted as portrait posing, which sets them apart from the other characters in the corpus for their lack of association with different types of labour. In contrast, when we look at work activities, such as sewing, selling, shoe mending, and making pulque, it is *Indios* and *Negros* who have the highest frequency of working depictions. In ratio terms, almost 40% of *Indios* and almost 50% of *Negros* are shown carrying out a work activity. These results indicate that the *Español* characters were represented as prosperous figures, though largely passive, especially when compared to *Indios*, *Negros*, *Mestizos*, and *Mulatos* who are shown being in charge of most of the basic economic activities.

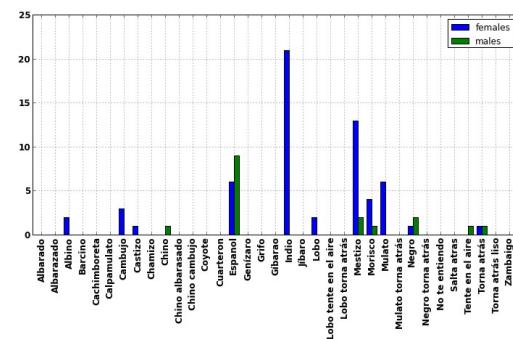


Fig. 6: Holding and carrying children by casta and gender frequency.

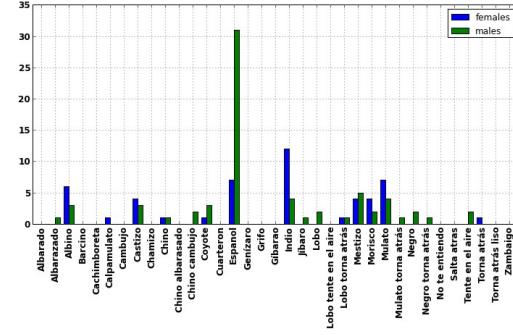


Fig. 7: Posing by casta frequency.

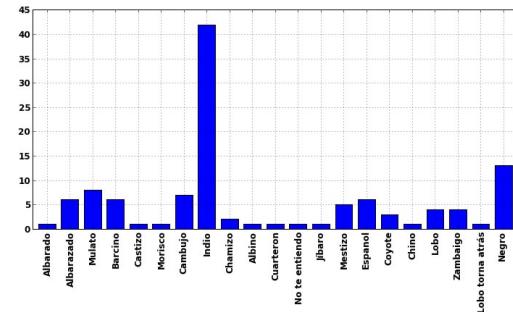


Fig. 8: Working by casta frequency.

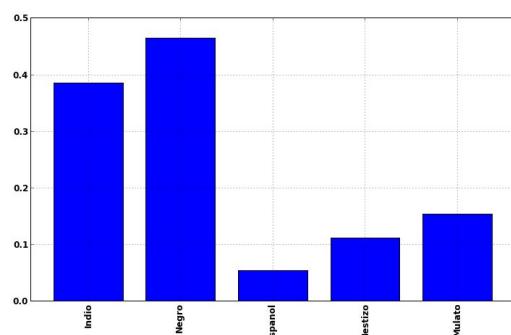


Fig. 9: Percentage by casta of working individuals.

4. Conclusions

By taking a general *casta* approach, this study argues that the notions of blood mending and Spanish prominence that informed and even produced *Casta* painting are fragile. These two principles of the caste system in place in New Spain at the time are consistent only as long as *castas* are seen individually. Conversely, when viewed as a group, *castas* not only outnumber the *Español*, *Indio* and *Negro* characters, they also signal that miscegenation was the standard. Furthermore, the contrast between the depiction of the *Españoles'* apparent prominence through non-laborious activities and the frequency of representations of *Indio*, *Negro*, *Mulato*, and *Mestizo* characters working, suggests an economic tension revealing a loss of Spanish relevance in the changing social landscape. Finally, *Casta* paintings not only fail at remedying the Spanish elite's anxieties, but, rather, stage the inevitability of miscegenation and even anticipate the demise of Spanish colonial rule.

References

1. Carrera, Magali. *Locating Race in Late Colonial Mexico*. Art Journal 57.3 (1998). Web. p. 38.
2. Katzew, Ilona. *Casta Painting: Images of Race in Eighteenth-century Mexico*. New Haven: Yale University Press, 2004. Print. p. 49.
3. Guzauskyte, Evelina. *Fragmented Borders, Fallen Men, Bestial Women: Violence in the Casta Paintings of Eighteenth-century New Spain*. Bulletin of Spanish Studies 86.2 (2009). Print. p. 176.
4. García Sáiz, María Concepción. *Las castas mexicanas: Un género de pictórico americano*. Milan: Olivetti, 1989. Print. p. 42.
5. Carrera, Magali. *Locating Race in Late Colonial Mexico*. Art Journal 57.3 (1998). Web. p. 45.
6. Léon, Nicolás. *Las castas*. Artes de Mexico 8 (1998). Print. p. 79.
7. De Orellana, Margarita. *La Fiebre De La Imagen En La Pintura De Castas*. Artes de Mexico 8 (1998). Print. p.58.
8. Katzew, Ilona. *Casta Painting: Identity and Social Stratification in Colonial Mexico*. New world orders: casta painting and colonial Latin America. University of Texas Press (1996). Print. p.13.
9. Katzew, Ilona. *Casta Painting: Images of Race in Eighteenth-century Mexico*. New Haven: Yale University Press, 2004. Print. p.49.
10. Paz, Octavio. *Sor Juana Inés de la Cruz, o, Las trampas de la fe*. Barcelona: Seix Barral, 1982. Print. p.42.

Matérialiser et rendre perceptible la transmission orale du savoir. L'édition électronique des cours d'Antoine Desgodets à l'Académie royale d'architecture en France, 1719-1728

Carvais, Robert

rcarvais@noos.fr

CNRS (France) – UMR 7074, Centre de théorie et analyse du droit

Chateau, Emmanuel

emchateau@laposte.net

Paris-Sorbonne (France), École nationale des chartes (France)

Introduction

Notre propos porte une réflexion sur la manière d'exposer et d'analyser une oralité. Dans le cadre d'un projet collectif, nous avons produit une édition critique numérique des cours d'Antoine Desgodets à l'Académie royale d'Architecture, en France, de 1719 à 1728¹. Avec cet enseignement, l'architecte opérait une rupture par rapport à ses prédecesseurs (François Blondel et La Hire père et fils) en proposant de mettre en ordre les savoirs d'une discipline professionnelle alors en formation, l'architecture. La particularité du matériau étudié réside dans le fait que les cours ne nous sont parvenus qu'à travers une tradition de manuscrits épars, dispersés dans plusieurs bibliothèques à travers le monde ; aucun des témoins ne semble être autographe. Les quatre thèmes abordés (les ordres, la commodité, les servitudes et le toisé) font l'objet d'une édition critique et ce sont au total trente trois manuscrits et textes imprimés qui ont été traités. En dehors de faire entrer des sources inédites dans l'histoire de l'architecture, des mathématiques et du droit, un des enjeux spécifiques du projet consistait à poser les jalons d'une réflexion sur la façon de traduire, et de mettre à disposition de la communauté des chercheurs, ce passage de l'oralité à l'écrit, puis dans certains cas à l'édition. Ici, tout en plaçant le projet dans le contexte des *digital humanities*, l'utilisation de méthodes d'édition électronique XML-TEI s'est révélée particulièrement adaptée². Ce sont les particularités de notre corpus protéiforme qui nous ont amenés à apporter notre contribution à cette problématique : d'abord l'instabilité du texte par le nombre des témoins conservés ; la place et le rôle des images dans le discours ; enfin la forte diffusion du texte juridique par ce que nous nommerons « l'édition à valeur ajoutée ».

1/- Traduire la labilité du corpus

Le corpus rassemblé contenait plusieurs états du prononcé de certains cours (ordres et toisés). De ce fait, notre but n'a pas été – même si nous en avons été tentés – de reconstituer une version unique de ces cours mais d'en conserver toute la fluidité. Les statuts si différents des témoins reproduits, sans que l'on sache précisément s'ils proviennent d'une dictée académique, d'une copie sur un texte maître disparu laissé à l'académie pour que les absents récupèrent le savoir ou de la simple copie par un clerc pour enrichir une bibliothèque privée ou par un praticien en quête de méthode, empêche de les traiter comme des textes identiques. L'oralité réside ici tant dans les traces de prononciation ou les problèmes d'écoute que dans la multiplicité des versions qui témoignent des différents prononcés du cours. Etant donné que ceux-ci sont dans leurs majeures parties semblables, mais jamais tout à fait identiques, seuls les moyens informatiques nous autorisent à ne pas figer définitivement ce savoir. Le jeu des variantes rendues en ligne propose une double lecture – à la fois rétrospective mais aussi prospective – des témoins sous l'œil averti du chercheur. La première fouille les détails pour requalifier le

témoin et se rapprocher de l'oralité originelle ; la seconde plus globale donne une idée d'un savoir voué à encadrer une nouvelle profession. Sans les spécificités de structuration et de précision sur les hésitations manuscrites offertes par les DH, et en particulier l'utilisation d'un balisage XML-TEI, ce matériau variant et composite serait resté insaisissable et intraitable par les procédés classiques de l'édition. L'alignement des témoins au sein du corpus a été déployé dans un fichier externe en utilisant la méthode de localisation référencée. La production d'artefacts numériques offre aux chercheurs l'opportunité de collecter les textes, de les analyser et de les mettre les uns en rapport avec les autres avec la souplesse inhérente au support numérique par l'intermédiaire d'une interface graphique centrée sur la comparaison.

2/- Dire les images au-delà de les montrer

L'architecture est une discipline pour laquelle le dessin n'est pas seulement primordial mais indispensable. Dans la plupart des pratiques architecturales, le dessin est le seul discours de l'architecture, le texte d'accompagnement ne sert souvent que d'explication des figures, voire de légende. Trois des cours de notre corpus contiennent des images, soit sous forme de planches comportant des plans ou des coupes et élévations (dans les ordres et les commodités), soit des détails décoratifs d'architecture, des croquis techniques de machines ou de figures géométriques (dans les toisés). Nous ne sommes pas face à des traités d'architecture imprimés, ouvrages savants classiques et référents, mais devant des cours pratiques prononcés devant un parterre d'élèves. Le texte ne fait alors qu'éclairer l'image, la commenter. Celle-ci devrait se suffire à elle-même. Le discours ne serait en quelque sorte que la forme orale de l'image exposée. L'édition critique numérique offre la possibilité de montrer les dessins dans leurs variantes et dans leur contexte, bref de les comparer et de les associer avec les textes correspondant par des mécanismes d'alignement et des modalités de consultation différentes, si tant est que nous soyons parvenus à lever les freins concernant les droits de diffusion. Un tel dispositif replace davantage les illustrations dans leur rôle pédagogique initial et les textes dans celui d'hypertexte, comme si nous avions dû les « dire ».

3/- La mise en abyme des discours

Pour les servitudes, nous n'avons découverts que deux manuscrits non pas du même cours mais préparatoires de celui-ci. Le premier est une compilation de sources, le second un commentaire d'une source principale. Or, d'après ce que nous en savons par ailleurs, le cours a été composé à partir du second complété par des éléments du premier. De surcroît, et c'est la particularité de ce cours de droit, il est le seul à avoir été publié de manière posthume. On peut dire que cette édition connaît un énorme succès si l'on sait que le texte a été commenté, mis à jour, complété, restructuré, annoté par plusieurs mains et compétences une trentaine de fois pendant un siècle. Les méthodes d'édition électronique nous ont permis de rendre compte et d'autoriser l'étude des continuateurs et suiveurs du texte par une sorte d'alignement successif des variantes (comparaison de textes, indexations). Il ne s'est pas agi de faire une édition génétique textuelle – désormais classique pour les œuvres littéraires –, à laquelle nous n'avons pas pu échapper d'ailleurs pour ce qui concerne la source coutumière, mais surtout de suivre les filiations d'idées dans le futur, bref d'avancer dans le temps plutôt que de reculer. La coutume est intrinsèquement le véhicule d'un contexte validé par sa diffusion sous forme imprimée, elle a sans cesse été glosée et revisitée. Nous aurions aimé pouvoir retrouver des manuscrits d'étudiants mais tous les textes rassemblés sont tous rattachés au commentaire de Desgodets. Le dispositif de consultation mis en place facilite l'analyse de tels enchaînements de textes afin de mettre au jour, selon les travaux de D. F. McKenzie, un « texte social » agrémenté d'un contexte historique, ce qui se révèle particulièrement riche pour des concepts juridiques qui évoluent lentement toujours en rapport avec leur environnement social, politique

et économique. C'est la constitution de ce « texte social » que nous voudrions maintenant pouvoir rendre explicite au moyen de visualisations.

Conclusion

Notre méthode numérique d'édition critique de cours appliquée à Desgodets mériterait d'être appliquée à d'autres matériaux inédits qui ont participé à transmettre du savoir. Le fait de ne pas s'arrêter à la publication d'un cours autographe mais à sa perception par les élèves et les praticiens, donne l'occasion de poser les bases d'une réflexion sur la normalisation de la discipline architecturale, la faisant passer d'un académisme engourdi à une solide professionnalisation. Le dispositif de consultation produit, qui est d'abord un dispositif d'étude, entend offrir au chercheur l'opportunité d'explorer avec fluidité toute la richesse de cette tradition de l'oral à l'écrit à travers le support numérique.

References

1. L'édition électronique des cours d'Antoine Desgodets sera accessible en ligne au début de l'année 2014 à l'adresse suivante : www.desgodets.net. Une équipe interdisciplinaire de quinze chercheurs a participé au projet qui a été financé par l'Agence nationale de la Recherche (ANR) de 2008 à 2013. Les membres de l'équipe étaient Joëlle Barreau, Basile Baudez, Anne Bondon, Robert Carvais, Pierre Caye, Emmanuel Château, Guillaume Fonkenell, Béatrice Gaillard, Juliette Hervu-Bélaud, Frédérique Lemerle, Olga Medvedkova, Linnéa Rollenhagen-Tilly, Hélène Roustéau-Chambon, Joël Sakarowitch, Werner Szambien et Dirk Van de Vijver.
2. Plusieurs projets pionniers en histoire de l'art ont constitué des modèles structurants (Blake Archive, Rossetti

Archive, etc.). Hormis la récente et excellente édition de la correspondance de Van Gogh, les projets d'édition électronique de sources primaires en histoire de l'art restent peu nombreux.

Computers and the Humanities, Image-Based Humanities Computing, vol. 36, n° 1, feb. 2002.

Burnard, L., O'Brien, K., O'Keeffe, Unsworth, J. (eds.) (2006), *Electronic Textual Editing*. New York: The Modern Language Association of America.

Hayles, K. (2004), "Print Is Flat, Code Is Deep: The Importance of Media-Specific Analysis". *Poetics Today*, 25, p. 67-90.

Hayles, K. (2003), "Translating Media: Why We Should Rethink Textuality", *The Yale Journal of Criticism*, 16(2), p. 263-90.

Kirschenbaum. M. (2001), "Materiality and Matter and Stuff: What Electronic Texts Are Made Of." *Electronic Book Review*, 1 Oct. 2001. <http://www.electronicbookreview.com/thread/electropoetics/>

McCarty, W. (2006), *Humanities Computing*. New York: Palgrave MacMillan.

McGann J. (2006), "From Text to Work: Digital Tools and the Emergence of the Social Text". *Romanticism on the Net*, n° 41-42, février-mai 2006. <http://id.erudit.org/iderudit/013153ar>.

McKenzie D. F. (1986), *Bibliography and the Sociology of Texts*. The Panizzi Lectures 1985. The British Library: London.

McKenzie D. F. (1962), *Making Meaning. "Printers of the Mind" and Other Essays*. Ed. Peter D. McDonald and Michael Suarez, SJ. U. of Massachusetts Press: Amherst and Boston, 2002.

McLuhan M., *The Gutenberg galaxy, the making of typographic man*. Toronto: University of Toronto Press.

Shillingsburg P. L. (1996), *Scholarly Editing in the Computer Age: Theory and Practice*, 3e édition. Ann Arbor: University of Michigan Press.

Shillingsburg P. L. (2006), *Gutenberg to Google: Electronic Representations of Literary Texts*. Cambridge University Press.

Waquet F. (2003), *Parler comme un livre. L'oralité et le savoir; XVIe-XXe siècles*. Paris : Albin Michel.

Waquet F. (2008), *Les enfants de Socrate, Filiations intellectuelles et transmission du savoir XVIIe-XIXe siècles*. Paris, Albin Michel.

Six terms fundamental to modelling transcription

Caton, Paul

King's College London, United Kingdom

Work begun by Huitfeldt and Sperberg-McQueen (2008) and continued jointly with Marcoux (2009, 2010) has given us a powerful and intuitive model of the abstract object T that counts as a successful transcription of an exemplar E. [1] (For convenience I will hereafter refer to these authors collectively as 'HSM' and their work as 'the HSM model'.)[2] Work on the HSM model is ongoing and a comprehensive formal account of the activity of transcription remains some way off, but extrapolating from what HSM have done so far we can begin to determine what is proper to a model of transcription per se, what is complementary to it, and what intra- and inter-model dependencies exist between entities. We can project, as it were, from the existing HSM model a bigger picture; but we stay always within the terms of the HSM model because it is the terms themselves we use for our projection. We thereby also clarify the terms themselves as they are used in the context of the model.

Here I focus on six terms fundamental to the 'bigger picture' and by elucidating these terms in relation to the model I sketch out the scope and composition of that picture: a necessary preliminary to any more detailed modelling to follow. The terms are: SURFACE, MARK, READING, TOKEN-SEQUENCE, EXEMPLAR, and DOCUMENT.

Consider the following scenario. A cave explorer discovers a new chamber and finds inside it three rock faces. One has complex patterns of scratches on it, each of the other two has painted lines on it. In all three cases what the explorer sees looks like writing, but she recognizes none of it. She takes a photograph of the rock face with the scratches and of one face with painted lines, then her camera fails. She pulls out a sketch pad and makes a faithful drawing of the other rock face with paint on it. Later she takes the photographs and drawing to an archeologist, who recognizes that the painted lines are indeed writing, in a language he knows, but that the forms of the characters are older and in many cases different from those in the current orthography of that language. He then makes written copies of the characters he sees in the photograph and drawing, this time using the current character forms to make them easier for other scholars to read. Doubtful about the scratched lines he informs a naturalist friend who visits the cave and confirms that all the scratches have been made by animals sharpening their claws on the rock face.

Now we retrace the conceptual movement of that narrative, introducing the fundamental terms in the appropriate places and using the context to clarify their relation to the HSM model and define them within that scope.

Each rock face is a SURFACE. A SURFACE is a thing: it is perceptible and measurable, and in its normal manner of existence can be returned to. The normal manner of existence of an electronic display, for example, is for a machine that generates it to be switched on and working properly: and while this is the case, we can return to the display.[3] A SURFACE is necessary for transcription; a SURFACE itself depends on nothing within the scope of this discussion. Of all things I describe here, a SURFACE is the closest to being a primitive entity.

Each scratch and painted line on the rock faces is a MARK. A MARK is a thing made upon a SURFACE by some agent. It is perceptible and measurable by contrast to the SURFACE, and therefore dependent upon the prior existence of the SURFACE.

The most complex of all the model-related entities is a READING. I will say it is normative - though not necessary - that READING is motivated. We are a communicative species and we actively look for instances of communication, willing to give the benefit of the doubt in many cases. MARKS are necessary for orthography, therefore the presence of MARKS implies the possible presence of writing. Consequently the presence of MARKS also implies the possible presence of text (here and throughout intended in the sense described in Caton 2013a), and text is written communication. Hence, in the normative case, an understanding of the possible presence of text motivates READING by an agent. In terms specific to the HSM model, READING is the process by which an agent attempts to discover and establish at least one TYPE-SEQUENCE in MARKS on a SURFACE by recognising certain MARKS to be certain TOKENS.

Because (in the normative case) READING is motivated, we must also grant that it may be entirely speculative. That is, it is acceptable that READING commences solely on the basis that MARKS are present on a SURFACE: there does not have to be certainty that at least one of those MARKS has token status.

Recall that our cave explorer could not assign token status to any of the marks she saw. Aware of her ignorance, she took steps (taking photographs and making a drawing) that have an interesting status in the overall picture because they seem to perform transcription without READING and therefore to deny that READING is necessary to transcription. But this is illusory. The goal of READING is to establish a TOKEN-SEQUENCE/TYPE-SEQUENCE, and the act of READING attempts this, assigning token status to MARKS where possible. There is no criterion of success for the activity: simply performing it is enough.[4] There doesn't have to be a specific TOKEN-SEQUENCE/TYPE-SEQUENCE at the end of it.

Instead of 'success', we distinguish three result-states of READING: **positive**, **negative**, and **zero**. A positive result-state means the agent assigns, with a greater-than-zero degree of certainty, token status to at least one MARK. A negative result-state means that the agent, with a greater-than-zero degree of certainty, decides no MARK has token status. A zero result-state means that for every MARK present the agent has

zero certainty that it is or is not a token. We shall return to this important point shortly.

An agent READS at least MARK by MARK (though more usually by groups of MARKS at a time), assigns token status where possible, and thereby constructs a TOKEN-SEQUENCE. A TOKEN-SEQUENCE cannot be empty: it must contain at least one TOKEN. Under any one READING R a TOKEN-SEQUENCE is neither 'right' nor 'wrong': it simply is the sequence under that READING, irrespective of the degree to which it corresponds to any text present on the SURFACE.

The dependence relation here is strictly one way and is of transcription upon the TOKEN-SEQUENCE produced by the READING. If no READING (minimal or informed) takes place, If there is a TOKEN-SEQUENCE, and if an agent desires to preserve its corresponding TYPE-SEQUENCE in another place by the activity of transcription, then that TOKEN-SEQUENCE assumes the role of EXEMPLAR with respect to the transcription, and in that respect we refer to it as the E-TOKEN-SEQUENCE. The manifestation in another place of the preserved TYPE-SEQUENCE as a TOKEN-SEQUENCE produces a TRANSCRIPTION (as result) and we refer to that sequence as the T-TOKEN-SEQUENCE.[5] Should an agent wish to transcribe this T-TOKEN-SEQUENCE, the sequence would then assume the role of EXEMPLAR with respect to this second transcription.

The essential insight of the HSM model is that transcription is that the T-TOKEN-SEQUENCE represents and preserves the E-TYPE-SEQUENCE. It should be clear then that with respect to the painted rock faces the cave explorer does perform transcription, despite her inability to consciously assign token status to any of the MARKS. By means of the photograph and the drawing, each E-TYPE-SEQUENCE of the painted MARKS is preserved and manifested in another place as a T-TOKEN-SEQUENCE, perceptible as MARKS on a SURFACE. Transcription is always possible when the READING result-state is either positive or zero. It does not have to happen deliberately, consciously - a transcription can be produced quite by accident. The archaeologist's transcription, unlike the explorer's, comes from positive result-state READING and he produces different T-TOKEN-SEQUENCES from the explorer, but they all represent the same E-TYPE-SEQUENCE.

Transcription is possible from a zero result-state READING, but only possible. The cave explorer's photograph of the scratches, for example, is not a transcription because no TOKEN-SEQUENCE is present on the scratched rock SURFACE and thus there is no E-TYPE-SEQUENCE to preserve.[6]

An act of transcription necessarily involves an EXEMPLAR, but does not necessarily involve a DOCUMENT. In relation to a model of transcription, and the model of READING which is necessary for it, we say that when there is at least one TOKEN-SEQUENCE / TYPE-SEQUENCE on a SURFACE, and the same READING that assigned token status to the MARKS also assigns TEXT status to the TOKEN-SEQUENCE / TYPE-SEQUENCE, then the SURFACE + TEXT combination acquires DOCUMENT status.[7] In a majority of cases an agent READS a SURFACE either certain it is a DOCUMENT or at least believing that highly likely. Hence DOCUMENT is a term frequently used in discussions of transcription, but transcription can take place without any DOCUMENT being present.

Notes

[1] For work responding to and building on the HSM model, see Caton 2009, 2013b.

[2] I must assume the reader's familiarity with the basics of the HSM model, in particular with their use of Peirce's concepts of token and type. Huitfeldt and Sperberg-McQueen 2008 gives the initial exposition; Caton 2013 provides a summary.

[3] I am avoiding the words 'material' and 'persistent' because (for the purposes of this discussion) those adjectives are not yet well enough defined with respect to the digital domain. Despite my expressive clumsiness I hope the reader understands that I am opposing the nature of SURFACE and MARK to the essentially transitory, of-the-moment nature of a phenomenon such as speech.

[4] Obviously this differs from the normal usage, where we expect someone performing the activity of reading to recognize a specific token sequence and consider their reading incorrect if they don't.

[5] Strictly speaking there is no T-TYPE-SEQUENCE prior to the existence of the T-TOKEN-SEQUENCE, only the E-TYPE-SEQUENCE. The T-TYPE-SEQUENCE is a product of the T-TOKEN-SEQUENCE.

[6] Because a negative result-state involves conscious judgement, it is entirely possible for one agent to perform two different READINGS with different result-states: one negative (by looking at MARKS on a SURFACE and deciding that none has token status) and one zero (by also taking a photograph of the MARKS).

[7] Because this ties DOCUMENT to a particular SURFACE it means every DOCUMENT is a unique object and not a 'repeatable symbolic expression' as discussed in Renear and Wickett 2009. I consider this uncontroversial as a constraint for the purposes of modelling, but I believe it also reflects a core sense of the common usage. Certainly 'document' is often used in an abstract sense, as in 'Magna Carta is an important document', but the signification is strongly tied to the idea of a particular piece of paper (or stone tablet, parchment scroll, email, etc.).

References

- Caton, Paul (2009). "Lost in Transcription: Types, Tokens, and Modality in Document Representation". Presented at Digital Humanities 2009, University of Maryland, College Park, June 2009.
- Caton, Paul (2013a). "On the term 'text' in digital humanities." Literary and Linguistic Computing 28 (2): 209-220. doi:10.1093/lrc/fqt001
- Caton, Paul (2013b). "Pure Transcriptional Markup". Presented at Digital Humanities 2013, University of Nebraska, Lincoln, July 2013.
- Huitfeldt, Claus and C. M. Sperberg-McQueen (2008). "What is transcription?" Literary and Linguistic Computing 23 (3): 295-310. doi:10.1093/lrc/fqn013
- Huitfeldt, Claus, Yves Marcoux and C. M. Sperberg-McQueen (2009). "What is transcription? (Part 2)." Presented at Digital Humanities 2009, University of Maryland, College Park, June 2009.
- Huitfeldt, Claus, Yves Marcoux and C. M. Sperberg-McQueen (2010). "Extension of the type/token distinction to document structure." Presented at Balisage: The Markup Conference 2010, Montréal, Canada, August 3 - 6, 2010. In Proceedings of Balisage: The Markup Conference 2010. Balisage Series on Markup Technologies, vol. 5 (2010). doi:10.4242/BalisageVol5.Huitfeldt01.
- Renear, Allen H., and Karen M. Wickett (2009). "Documents Cannot Be Edited." Presented at Balisage: The Markup Conference 2009, Montréal, Canada, August 11 - 14, 2009. In Proceedings of Balisage: The Markup Conference 2009. Balisage Series on Markup Technologies, vol. 3 (2009). doi:10.4242/BalisageVol3.Renear01.

Z-Axis Scholarship: Modeling How Modernists Write the City

Christie, Alex

Ross, Stephen
University of Victoria

Sayers, Jentery
University of Victoria

Tanigawa, Katie
University of Victoria

INKE-MVP Research Team
University of Victoria

Introduction

"Humanities on the Z-axis" is an interdisciplinary research project that works across modernist studies, geospatial humanities, and desktop fabrication. Through a combination of techniques in three-dimensional (3D) fabrication, geospatial mapping, speculative computing, and pattern analysis, z-axis research expresses the geospatial narratives of modernist novels by geo-referencing them and then using that geo-data to transform base layers of maps from the modernist period. The output of the research includes warped, 3D maps of cities (e.g., Paris and Dublin) central to modernist literary production. These maps can be viewed as 3D models on a screen or as physical prototypes in hand, and they are currently being transformed using geo-data drawn from novels by Djuna Barnes, James Joyce, and Jean Rhys. Ultimately, they show how modernist authors wrote the city, and findings suggest they contradict existing research in modernist studies about how, exactly, cities are expressed in modernist novels.

Research Problems

This project addresses two specific research problems that currently exist across geospatial humanities and modernist studies. First, in fields such as digital humanities, geospatial mapping techniques (Moretti 2005) and data visualizations (Bostock 2012) tend to produce isomorphic cartographies or flat representations of data (Drucker 2011), even when literature resists this type of representation. One consequence of these flat or isomorphic approaches is that they tend to ignore the importance of subjective experience to literary criticism (in particular) and the humanities (in general). Second, and related to the first point, mapping techniques in geospatial humanities research can too easily be applied across literary periods without regard to the historical, material, or formal differences between texts, especially when something like Google Maps or Google Earth is the core technology. As a result, geospatial methodologies do not always persuasively correspond with the literary period, aesthetics, and textual particulars under examination.

In response to these problems, z-axis research tailors mapping practices to suit the needs of literary periods. For instance, modernist literature deliberately resists isomorphic representations of geographic space (Vidler 2000). It also frequently treats the city as a medium, which is represented through fiction. Consequently, we argue that modernism not only calls for speculative, non-isomorphic modes of geographical expression (i.e., deformed maps) but also techniques that engage geographic representation directly (i.e., by distorting a map's base layer instead of "pinning" data on top of it). Additionally, the z-axis methodology involves a "text-first" workflow wherein the specificities of the text practically dictate the mapping method and, by extension, the aesthetics of the transformed, 3D maps.

Research Questions

Z-axis research currently asks the following research questions of geospatial humanities and modernist studies:

- How and to what extent do geospatial approaches to modernist novels benefit from distinct methods of analysis? With what implications on existing geospatial methods in digital humanities?
- How do modernist authors write the modernist city? Through 3D maps, how can we compare multiple, literary versions of the same modernist city, using the same base map?
- How can traditions in speculative computing (Drucker and Nowviskie 2004; Drucker 2009) and deformation (Samuels and McGann 1999) be applied to geospatial analysis and the modernist novel (Nowviskie et al. 2013)?
- In modeling and fabrication practices, to what degree (if at all) do 3D-printed maps afford interpretations that screen-based

maps do not? Where visual expression is concerned, how do we put screen and print into conversation, and to what effects on the trajectories of scholarly communication?

Literature Review

Many geospatial projects in digital humanities are largely two-dimensional and rely significantly on isomorphism. In the case of modernist cities, projects such as Walking Ulysses (2012) and WatsonWalk (2012) pin events to flat base maps. While the Scholars' Lab at the University of Virginia is working on "social and spatial maps of modernist correspondence," projects of this sort are rare in the field. Building on these initiatives, z-axis research uses historical maps as a medium for expressing social and cultural currents in the modernist city.

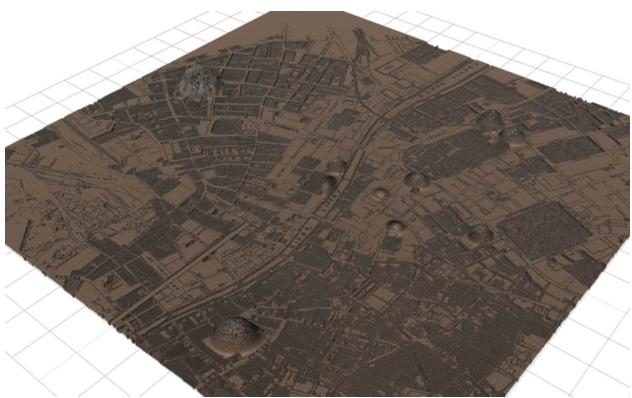
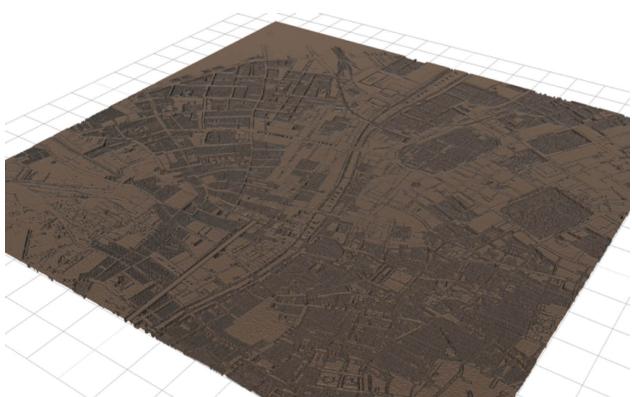
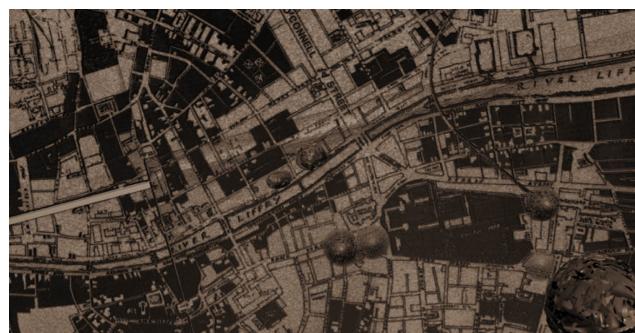
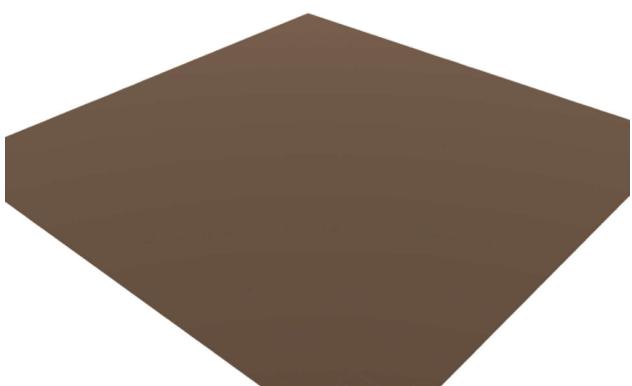
At the same time, many writers interested in modernity have documented the constructed character of modernist geographies. Henri Lefebvre (1974) unpacks the social production of urban space, arguing that discrete social and spatial practices are embedded in different cities. Elsewhere, Foucault's heterotopias (1984) and Marc Augé's (1992) non-spaces chart the social construction of space as it ruptures geographic locales, producing overlapping and contradictory spaces. Embedding the social and political nature of modern cities into their narrative, many modernist novels construct the city or treat it as a medium. In his work on cartographical rhetoric in Ulysses, Jon Heggland argues that the very act of mapping within the text is seen as a way of knowing. Richard Zeikowitz (2005) and Deborah Parsons (2000) similarly analyze the construction of feminist cityscapes in Jean Rhys's works, while Amy Wells-Lynn (2005) reveals the way Djuna Barnes and other female modernists "construct new Parisian geographic and literary female spaces" (79).

When combined, geospatial research in both digital humanities and modernist studies suggests that specific mapping approaches to literary modernism underscore the importance of socially constructed, geographic space. Here, work in speculative computing (Drucker and Nowviskie 2004; Drucker 2009) and deformation (Samuels and McGann 1999) provides precedent when blending computation with humanities inquiry.

Method and Workflow

When studying modernist novels, the current workflow for z-axis research is bifurcated into two processes. The first process involves geo-referencing plain-text versions of modernist novels, and includes the following steps: 1) use VueScan software to produce high-resolution scans (600 dpi) of the text; 2) run the text through ABBYY FineReader to render it machine-readable; 3) where necessary, correct the text; 4) geolocate the narrative of the text in XML (if a character appears in a certain location while imagining or talking about another location, then the location is tagged as the place where the narrative occurs); 5) conduct a word-count to see how many words are nested in certain geographic locations; 6) record these numbers in a spreadsheet; and 7) divide the number of words per location by the total number of words in the text to produce a ratio.

The second process mobilizes the geo-data and ratio from the first process to transform historical maps in 3D. It includes the following steps: 1) use a high-quality scanner to digitize an archival map; 2) use Photoshop to convert the scanned archival map into a displacement map; 3) apply the displacement map to a flat subdivided mesh using Autodesk's Mudbox, then scale the plane along its z-axis to render details from the displacement map as changes in elevation (figures 1-3); 4) procedurally apply the bulge function for each area indicated by the data model, warping the three-dimensional map along its z-axis (figures 4, 5), with differences in warping determined by the geo-data ratio; 5) use MeshLab to determine if the Mudbox model is watertight; if it is not, then automatically correct errors if they are not glaring; and 6) print the model using a desktop 3D printer.



Findings

One of the key findings of this research is the articulation of a methodology and workflow for expressing the geospatial narratives of modernist novels through transformance and speculative computing. Additional, related findings suggest that, contrary to an abundance of modernist scholarship (e.g., Hegglund 2003), James Joyce's *Ulysses* does not provide an isomorphic representation of Dublin. Instead, the novel presents a biased version of the city, privileging specific geographic areas over others. What's more, in the case of modernist novels about Paris, constructions of the city are highly contingent upon constructions of sexuality, especially when Barnes's Paris is compared with that of Rhys. That is, the sexual politics of literary Paris dramatically influence how and what parts of it are represented in texts from the 1920s and '30s. Finally, where comparisons between screen and print media are concerned, findings suggest that the latter not only affords tactile engagements lacking in the former but also bypass many visual design problems, including tendencies to squeeze too much complex information into a single frame or window. More generally, our findings suggest that the humanities can be empowered through the material transformation of scholarly communication, including evocative objects and publications in 3D.

Acknowledgements

This research has been conducted at the Electronic Textual Cultures Lab and the Maker Lab in the Humanities at the University of Victoria with support from the Modernist Versions Project (MVP) and Implementing New Knowledge Environments (INKE). The research is supported by the Social Sciences and Humanities Research Council (SSHRC).

References

- Augé, Marc** (1997). *Non-Lieux: introduction à une anthropologie de la surmodernité*. Paris: Le Seuil.
- Barnes, Djuna** (1995). *Nightwood: The Original Version and Related Drafts*. Ed. Plumb, Cheryl J. Normal, IL: Dalkey Archive Press.
- Bostock, Mike** (2012). "D3.js: Data-Driven Documents." d3js.org .
- Declan, Kiberd** (2009). *Ulysses and Us: The Art of Everyday Living*. London: Faber and Faber.
- Drucker, Johanna** (2009). *SpecLab: Digital Aesthetics and Projects In Speculative Computing*. Chicago: Chicago University Press.
- . "Humanities Approaches to Graphical Display." (2011) Digital Humanities Quarterly. 5.1.
- Drucker, Johanna and Bethany Nowviskie**. "Speculative Computing: Aesthetic provocation in Humanities Computing." A Companion to Digital Humanities (2004). ed. Susan Schreibman, Ray Siemens, John Unsworth. Oxford: Blackwell.
- Foucault, Michel and Jay Miskowiec** (1986). "Of Other Spaces." Diacritics 16.1 (Spring 1986). 22-27.
- Joyce, James** (1922). *Ulysses*. Paris: Shakespeare & Co.

- Kraus, Kari.** (2009) "Conjectural Criticism: Computing Past and Future Texts." *Digital Humanities Quarterly* 3.4 (Fall 2009): n. pag.
- Hegglund, Jon** (2003. "Ulysses and the Rhetoric of Cartography." *Twentieth Century Literature*. 49.2. 164-192.
- Lefebvre, Henri** (1991. *The Production of Space*. Trans. Donald Nicholson-Smith. Oxford: Blackwell.
- Moretti, Franco** (2005). *Graphs, maps, trees: abstract models for a literary history*. London: Verso.
- Nowviskie, Bethany**. *Neatline*. LLC.
- Parsons, Deborah** (2000). *Street Walking the Metropolis: Women, the City, and Modernity*. Oxford: Oxford UP.
- Ramsay, Stephen and Geoffrey Rockwell** (2012). "Developing Things: Notes toward an Epistemology of Building in the Digital Humanities." *Debates in the Digital Humanities*. Ed. Matthew K. Gold. Minneapolis: University of Minnesota Press. 75-84.
- Rhys, Jean**. *Good Morning, Midnight* (1970). New York: Harper & Row. First published in 1939.
- Rhys, Jean** (1969). *Quartet*. London: Andre Deutsch. First published in 1928.
- Seidel, Michael** (1976). *Epic Geography: James Joyce's Ulysses*. Princeton UP: 1976.
- Vidler, Anthony** (2000). *Warped Space*. Cambridge, Mass: MIT Press.
- Wells-Lynn, Amy** (2005). "The Intertextual, Sexually-Coded Rue Jacob: A Geocritical Approach to Djuna Barnes, Natalie Barney, and Radclyffe Hall." *South Central Review* 22.3 (Fall 2005). 78-112.
- Zeikowitz, Richard** (2005). "Writing a Feminine Paris in Jean Rhys's 'Quartet.'" *Journal of Modern Literature*. 28:2. 1-17.

Book History and Software Tools: Examining Typefaces for OCR Training in eMOP

Christy, Matthew

mchristy@tamu.edu
Texas A&M University

Samuelson, Todd

toddsamuelson@library.tamu.edu
Texas A&M University

Torabi, Katayoun

torabik@neo.tamu.edu
Texas A&M University

Tarpley, Bryan

bptarpley@neo.tamu.edu
Texas A&M University

Grumbach, Elizabeth

egrumbac@tamu.edu
Texas A&M University

In 1936, the notable English bibliographer A. W. Pollard admitted in his Preface to Frank Isaac's English Printers' Types of the Sixteenth Century that "[he] had a very poor eye for distinguishing types and a very poor head for remembering them."¹ Pollard is hardly alone among experts in the history of printing in this shortcoming. Even among scholars with decades of experience in scrutinizing features of the printed book, the ability to distinguish and identify typefaces is a notorious challenge. The literature about early type designs and designers (known as punchcutters) is partial and contradictory; the variations in typefaces are subtle and, at times, inconclusive; and the ability to make differentiations has been considered less a matter of regimented principle than of elusive skill. As Harry Carter suggested, "it is evident that in considering the face of a fount of type we are in a world of art, . . . not a mechanical proceeding or anything susceptible of scientific treatment."²

However, it is precisely the consideration of "founts of type" that is currently engaging a majority of the Early Modern OCR

Project (eMOP) team. eMOP, a 2-year Mellon Foundation-funded grant project underway at the Initiative for Digital Humanities, Media, and Culture (IDHMC) at Texas A&M University, aims to OCR the documents that comprise the Eighteenth Century Collections Online (ECCO) and Early English Books Online (EEBO) collections. As a project that involves collecting and aggregating huge amounts of data, OCR'ing 45 million page images on a high-performance computing cluster, and the development of several software tools and services, eMOP is technology-laden. But at its heart eMOP is a Humanities project, conceived by Humanists, driven by the needs of Humanities scholars, and supported throughout by book history and an understanding of the development of print type in the 15th-18th Centuries.

For eMOP's book historian, Dr. Todd Samuelson, one of the difficulties in conceptualizing the scope of eMOP has centered in a potential conflict between DH methodology (as encompassed by "big data") and the traditional means of approaching type identification: as Carter noted, it is an art steeped in years of hard-won practice rather than a science with predictable and reproducible models. While DH is focused on humanities questions and methodologies, it does employ scientific principles as well, especially when dealing with a very large set of documents, and conflict can arise by trying to synthesize a skill set based on minutiae with an extremely large data set. By contrast, even when big data projects incorporate crowdsourcing and the oversight of human experts, they require the ability to find readily transferrable commonalities, rather than to establish proficiency in a small number of experts. In the course of the eMOP project, we have found that the development or adoption of specific software tools has helped to ameliorate this conflict and incorporate type history scholarship into the training of OCR engines.

One of the ideas driving eMOP work is that, by training OCR engines to recognize specific early modern fonts, we can increase the accuracy of those engines when used to OCR documents printed in those fonts. To accomplish this, the eMOP team has spent most of the last year investigating font history, creating a database of early modern printers and the fonts they used, and developing and testing tools and techniques to train Tesseract (an open-source OCR engine) to recognize these fonts. The ability to distinguish between different, but sometimes closely related, fonts, and to train Tesseract to recognize these distinctions has been a central focus. For example, the general classification of different families of typefaces has been attempted by book historians, including Adrian Weiss, who categorized unknown English typefaces of a certain period as either "S-face" or "Y-face".³ So, though the source of the typeface may not be ascertainable, certain characteristics can be defined which allow scholars (and potentially OCR engines) to identify and group the typefaces more accurately.

As has already been noted, identifying examples of S- and Y-face characters and distinguishing between them, especially when both can be present in one document, is a difficult enough task for an expert. Trying to find all instances of the lower-case letter 'w' in a document, as an example, and then deciding which exemplars match some specified "ideal" is difficult and time consuming. Fortunately, eMOP has software tools that can drastically simplify this task, and even allow non-experts to do some of the work. Those tools were originally developed to create training for Tesseract to recognize early modern typefaces, but can also be applied to support research into the typefaces themselves.

To create specific font training for the Tesseract OCR engine, a team of undergraduate student workers, lead by IDHMC graduate student Katayoun Torabi, first process the available page images using Aletheia Desktop. (Aletheia was developed by the Pattern Recognition and Image Analysis (PRImA) Research Laboratory at the University of Salford. Apostolos Antonacopoulos, IMPACT Work Package leader for PRImA, University of Salford, has made Aletheia and other tools available at <http://www.primaresearch.org/tools.php>.) Aletheia Desktop includes several semi-automated tools that identify and define layout regions, lines, words, and individual characters (glyphs) within documents. Aletheia reads the text in

the page image (using Tesseract) and assigns a Unicode value for each letter, number, and punctuation mark.

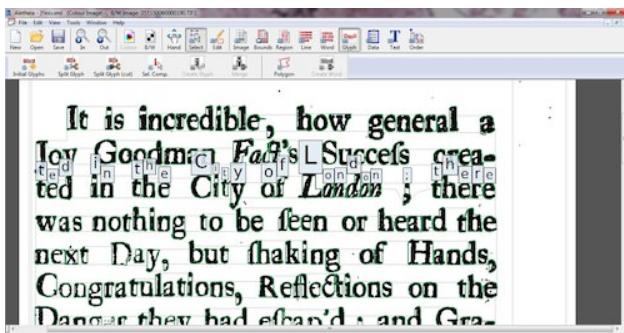


Fig. 1: Aletheia Desktop with identified glyphs and some of their associated Unicode values.

As output, Aletheia creates an XML file that contains a set of XY coordinates, along with the associated Unicode value, for each identified glyph. The data contained in this XML file is then ingested or imported into a tool created by IDHMC graduate student Bryan Tarpley called Franken+. Franken+ uses a MySQL database to associate each glyph image with its corresponding Unicode character. The user can then select any glyph from a drop down menu to see every instance of that character in a window (Fig. 2). With every instance of a particular glyph (for example all the 'a's) from a document available in one window, the user can quickly identify mislabeled glyphs and choose the best exemplar (or exemplars) for each glyph in that font set (Fig. 2). Once the user has isolated the best instance(s) of each character, Franken+ uses a standard text document to produce a set of synthetic TIFF images and XML files, producing a "Franken-text" with only these ideal characters. This Franken-text matches the characteristics of Tesseract's expected training file and so can be used to train Tesseract to recognize the typeface being processed.



Fig. 2: Some images of the Frank+ user interface.

eMOP's book history team immediately realized that the capabilities of Aletheia and Franken+ would tremendously benefit their research into the S-face vs Y-face font question. The ability of Franken+ to display all instances of a given letter from a set of page images in one window dramatically simplifies the task of identifying all examples of any letter in a set of pages. And, being able to examine all these examples alongside each other makes comparing similarities or differences much easier and faster (Fig. 3). After a quick installation of Franken+ and less than an hour of training, the book history team was able to commence work on their research question in earnest. Since Franken+ was introduced at the 2013 Doc Eng Conference,⁴ the eMOP team has been contacted by several international scholars interested in learning more about Franken+ for use in their research on typefaces.



Fig. 3: An image from Franken+ of a set of exemplars of the "a" glyph from one document.

The study of early modern fonts is a road less traveled in the landscape of Humanities research. Based as it is on minutiae and requiring incredible attention to detail, this work traditionally has been left to a handful of individual scholars. However, the development of Franken+ for eMOP, when used in conjunction with Aletheia, promises to open up this field of study to scholars who may have been interested in it, but found the challenges too daunting. This paper will describe aspects of the eMOP work being done in the field of early modern type research, and will introduce Franken+ as a valuable new tool in this research. The creation of tools like Franken+ have the potential to increase attention and alter research methodologies for this field.

References

- Pollard, A. W.** (1936) *Preface*. Isaac, Frank. English Printers' Types of the Sixteenth Century. Oxford: Oxford UP. (vii).
- Carter, Harry Graham** (1969). *A View of Early Typography Up to About 1600*. Oxford: Clarendon.
- Weiss, Adrian.** (1990) *Font Analysis as a Bibliographical Method: the Elizabethan Play-Quarto Printers and Compositors*. Studies in Bibliography 43: 95-164.
- Torabi, Katayoun, Jessica Durgan, and Bryan Tarpley** (2013). *Early Modern OCR Project (eMOP) at Texas A&M University: Using Aletheia to Train Tesseract*. ACM Press. 23. doi:10.1145/2494266.2494304. Web. 31 Oct. 2013.

Diagnosing Page Image Problems with Post-OCR Triage for eMOP

Christy, Matthew
mchristy@tamu.edu
Texas A&M University

Auvil, Loretta
lauvil@illinois.edu
University of Illinois

Gutierrez-Osuna, Ricardo
rgutier@cse.tamu.edu
Texas A&M University

Capitanu, Boris
capitanu@illinois.edu
University of Illinois

Gupta, Anshul
ag@guptaanshul.com
Texas A&M University

Grumbach, Elizabeth
egrumbac@tamu.edu
Texas A&M University

The Early Modern OCR Project (eMOP), currently underway at the Initiative for Digital Humanities, Media, and Culture (IDHMC) at Texas A&M University, is a Mellon Foundation-funded endeavor tasked with improving, or creating, OCR (optical character recognition) for the Eighteenth Century Collections Online (ECCO) and Early English Books Online (EEBO) collections. The basic premise of eMOP is to 1) use book history to identify the fonts represented in the collections and the printers that used them; 2) train open source OCR engines on those fonts; and 3) OCR documents using an engine trained on the font specific to each documents. In addition, as a Mellon Foundation-funded project eMOP is tasked with using open-source solutions and producing open-source tools, workflows, and processes that are reproducible and which can be implemented by other scholars in their own digitization projects. One of eMOP's end products will be an open-source workflow of our entire process using Taverna.

As eMOP enters its second year, intensive work on developing and testing training for the Tesseract OCR engine has demonstrated a failing in the three-fold basic premise. Many of the page images which we are trying to OCR are of such poor quality that no amount of training will produce OCR results that meet the standards we have set for the grant outcome.¹ These images are already binarized, low-quality, low-resolution, digitized images of microfilm, converted from photographs—4 decades and 3 media generations removed from the originals. Typical problems include noise, bleedthrough, skewing, and warping, but there are many more. There already exist many algorithms that can fix most of the problems extant in our collection of page images.^{2 3} Applied during a pre-processing stage, these algorithms have the potential to improve page image quality to the point that they can yield excellent OCR results. But with approximately 45 million pages in eMOP's data set, determining which pages need which kind of pre-processing proved problematic at best.

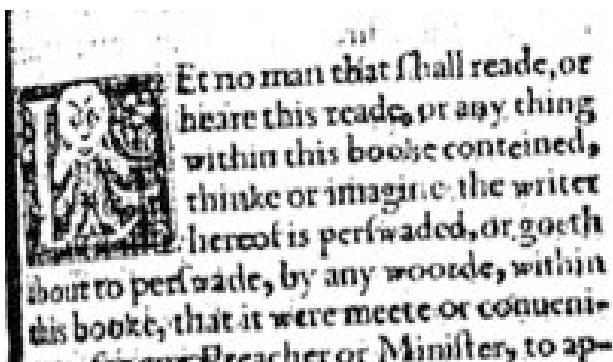


Fig. 1: A sample of part of a page image from the eMOP collection showing skew, noise, bleedthrough, over-inking, and an image.

To this end, the eMOP management team, along with our collaborators, Loretta Avil and Boris Capitanu at SEASR (Software Environment for the Advancement of Scholarly Research, at the University of Illinois, Urbana-Champaign), and Dr. Ricardo Gutierrez-Osuna and graduate student Anshul Gupta of Texas A&M University, decided to focus our proposed post-processing triage workflow on the problems that exist in our page image inputs. Originally stated, our triage process would examine OCR results and decide whether the documents would be routed to different tools being built for eMOP to perform automatic word correction, crowd-sourced line segmentation correction, by-hand font identification, or automated re-OCRing with different font training. However, the presence of so many low quality page images in our input required a more robust system for handling the output. What we needed was a triage process that would allow us to programmatically diagnose our input documents based on the output of our OCR system.

The open-source Tesseract OCR engine is capable of producing both plain text files and files in an XML-like format called hOCR. hOCR files contain wrappers around each found word, line, paragraph, and region, and these wrappers contain bounding box coordinates for each entity (Fig. 2). A close

examination of the text and hOCR results for nearly 600 poor quality page images revealed certain patterns, or 'cues', which could be used, singly or in combination, to uniquely predict individual problems that exist in the original page images.

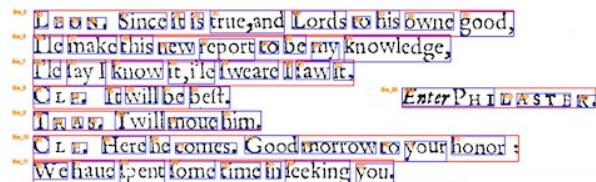


Fig. 2: Bounding boxes for lines (red) and words (blue) drawn on a page images based on hOCR output.

For example, documents printed in a blackletter or gothic font, but OCR'd with Tesseract trained for a roman font produce a text file with a character frequency distribution different from that expected of English language documents. Basically, if Tesseract is trained with a roman font, characters printed in a blackletter font look predominantly like m, n, u, and l. Similarly, documents containing a lot of noise (e.g. numerous spots and blotches on the page) typically produce "words" found in areas of the page outside of the main text area, have word bounding boxes of widely varying heights, and have line bounding boxes that overlap. Page images that exhibit heavy skewing (the text lines are tilted at an angle from the horizontal) also pose problems for Tesseract, as it will often begin reading one line and then at some point jump to the line above or below (depending on the direction of the skew) to finish reading the "line." In these cases the hOCR again contains overlapping line bounding boxes, but also has word bounding boxes that don't have contiguous coordinates, i.e. it is finding words out of the reading order as they appear on the page to a human reader. These are just a few examples that demonstrate the problems we've encountered and the cues we've discovered to identify them, which we will identify in this paper.

Cues like these and others have provided us with the mechanism we were looking for to identify page image problems based on OCR output. In order to take full advantage of this information however, we are also developing a full post-processing workflow. Beginning with OCR results, the output of this workflow will be either 95%, or better, corrected text or a per-page indicator describing what kind of pre-processing should be performed before each page is re-OCR'd.

We are also working with our collaborators on developing a mechanism to assess the quality of our OCR output. We have combined different analysis techniques developed by collaborators at SEASR and Texas A&M University, to examine text data (examining character unigram frequency distributions and word lengths), page data, (determining the main text area of the page and looking for outliers), and hOCR bounding boxes (calculating box heights and widths). Applying these mechanisms to the results of each page will yield a score that constitutes a prediction of how the document would compare to a ground-truth transcription. Test results show a strong correlation between these predicted scores and actual scores produced on documents that do have ground-truth available.

Page results receiving a high enough score can then be sent for further text analysis, including dictionary look-ups, to correct as much of the OCR output as possible. Those pages that receive scores below the threshold undergo an iterative process of looking for different cues in order to identify the likely reason the OCR process failed for each page.

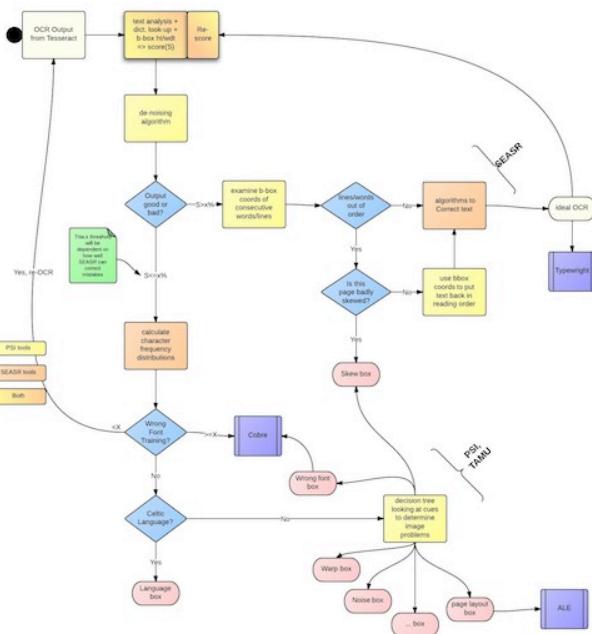


Fig. 3: Proposed eMOP post-processing workflow.

Much work has already been done with regard to OCR post-processing, but it has concentrated on questions of identifying and correcting bad OCR.⁴ ⁵ In this paper we will report on the development of an OCR post-processing workflow that can evaluate and identify a broad range of defects common to page images of early modern printed documents. The result of this workflow can then be funneled into a pre-processing and re-OCR'ing process later. We plan, by grant-end, to release an open-source workflow and code that can be used by other groups or individuals engaging in large-scale OCR projects. Given the inherent problems that these documents pose for OCR engines, we view this kind of analysis as a vital step forward in the comprehensive understanding and digitization of large collections of early modern printed documents.

References

1. Singh, Chandan, Nitin Bhatia, and Amandeep Kaur. *Hough Transform Based Fast Skew Detection and Accurate Skew Correction Methods*. *Pattern Recognition* 41.12 (2008): 3528–3546. CrossRef. Web. 30 Oct. 2013.
2. Subramaniam, L. Venkata et al. *A Survey of Types of Text Noise and Techniques to Handle Noisy Text*. ACM Press, 2009. 115. doi:10.1016/j.patcog.2008.06.002. Web. 30 Oct. 2013.
3. Taghva, Kazem, Thomas Nartker, and Julie Borsack. *Information Access in the Presence of OCR Errors*. ACM Press, 2004. 1–8. doi:10.1145/1031442.1031443. Web. 30 Oct. 2013.
4. Wudtke, Richard, Christoph Ringlstetter, and Klaus U. Schulz. *Recognizing Garbage in OCR Output on Historical Documents*. ACM Press, 2011. 1. doi:10.1145/2034617.2034626. Web. 30 Oct. 2013.
5. Borovikov, Eugene, Ilya Zavorin, and Mark Turner. *A Filter Based post-OCR Accuracy Boost System*. ACM Press, 2004. 23–28. doi:10.1145/1031442.1031446. Web. 30 Oct. 2013.

Developing for Distant Listening: Developing Computational Tools for Sound Analysis By Framing User Requirements within Critical Theories for Sound Studies

Clement, Tanya

University of Texas at Austin, United States of America

Developing for Distant Listening: Developing Computational Tools for Sound Analysis By Framing User Requirements within Critical Theories for Sound Studies

The Council on Library and Information Resources (CLIR) and the Library of Congress (LoC) issued a 2010 report that suggests that if we do not use sound archives, our cultural heritage institutions will not preserve them. Nancy Davenport, previous president of CLIR, concludes that users want unfettered access and better discovery tools for what she calls “deep listening” (what Charles Bernstein calls “close listening”) or “listening for content, in note, performance, mood, texture, and technology.” It is a typical digital humanities problem: Without a better understanding of what such listening entails, we cannot build tools that afford such listening; and, because we lack the tools, humanists struggle to imagine how to describe the access they want -- what Jerome McGann calls “imagining what you don’t know.” In an attempt to imagine how to facilitate distant listening with computation, this paper positions user requirements for critical listening software within the context of critical listening theories.

Critical Listening Theories

Walter J. Ong once announced that recording technologies have heralded a new age in the study of the “voice, muted by script and print.” Humanists hold a range of theories and perspectives on how to study music or sound aesthetics in experimental poetry or how to contextualize sounds of the recording space (such as the whir of an old air conditioning or babies crying in the background) or the recording machine (such as the clicks and pops and pauses).

In particular, theoretical perspectives on “the voice” are useful in identifying the role sonic features that are discoverable by computation can play while close listening. Most such theories, for example, position sonic vocal traits as meaningful only within the context of a structural code for meaning such as language. Roland Barthes identifies two aspects of the voice in vocal music, for instance, that contribute to meaning making: the pheno-song, which refers to the structured elements of a piece such as speech or melody (or language codes) and the geno-song, which is the material or corporal aspect of the voice, the “volume of the singing and speaking voice, the space where significations germinate” (or sonic features). Privileging the pheno-song as more productive for communicating meaning, Barthes maintains that the geno-song –having “nothing to do with communication, representation (of feelings), expression”—is a system for transmitting that meaning.

Similarly, Michael Chion asserts that sonic features have meaning but that it is our lack of a descriptive system or hermeneutics that precludes our ability to make sense of these features. Chion approaches sound study by parsing listening into causal (for the source of the sound), semantic (to interpret a message), and reduced (to identify sonic traits) listening. Chion argues that reduced listening precludes meaning making for two reasons: (1) the “fixity” of sonic features required for close listening to sonic traits makes sound “physical data” that do not represent what was actually spoken or actually heard in real time “presence”; and (2) our language for describing how we make meaning with such traits is “totally inadequate.” Consequently, because of issues of fixity and inadequate identifiers, reduced listening is “an enterprise that is new, fruitful, and hardly natural.”

This argument, however, that the voice is only meaningful in the context of speech that transmits a message is a logocentric theoretical stance that has been readily contested. Adriana Cavarero who seeks to “understand speech from the perspective of the voice instead of from the perspective of language” wants to “pull speech itself from the deadly grip of logocentrism.” Caravero critiques the viewpoint of

scholars such as Walter Ong and Marshall MacLuhan who at once essentialize the voice as “presence” and disembody and mythicize orality. Similarly, Mladen Dolar considers a “linguistics of non-voices” including coughing, hiccups, babbling, screaming, laughing, and singing, placing these sounds outside of the phonemic structure yet not outside of the linguistic structure. He argues that “It is not that our vocabulary is scanty and its deficiency should be remedied: faced with the voice, words structurally fail.” Finding possibilities for study in aspects of the voice such as accent, intonation, and timbre, Dolar asks the question at the heart of all of these queries: “how can we pursue this dimension of the voice?”

User Needs

User perspectives on the kinds of access and analysis advanced technologies with sound can facilitate were gathered by Clement as part of the HiPSTAS (High Performance Sound Technologies for Access and Scholarship) project. HiPSTAS is an NEH-funded, year-long Institute for Advanced Topics in the Humanities for librarians, information scientists, and humanities scholars who work with spoken word collections. Such collections include PennSound’s poetry archive, the American Folklife Center of the Library of Congress, oral histories in StoryCorps, and recordings from more than 50 tribes across Native America in the American Philosophical Society’s Native American Collection among other collections of interest to the participants. User perspectives were gathered from the 20 participants in three ways: (1) through Institute applications; (2) through pre-Institute interviews; and (3) through post-Institute surveys and project reports.

This data shows that defining the sonic features that map to specific cultural characteristics of “the voice” in spoken word recordings was not how participants phrased their research interests. One participant, for example, who was interested in working with the PennSound archive, wanted to consider “media ecologies” by analyzing “sounded affinities between poets” or “concepts of community poetics through sound;” this participant wanted “to look at groups of poets who have a common locale in terms of their community formation” and to use these clusterings to investigate how software “may or may not track affinities across gender lines.” Another participant analyzing PennSound was interested in “identifying, exploring, and categorizing performance variants of the same texts” such as “an aural/visual equivalent to the Versioning Machine” and enabling “a kind of distant listening, flagging and visualizing generic features of poetry performance traditions (Is there, for example, a New York School style of oral delivery?)” Other concerns were focused on reorienting how the archive is discoverable by enabling a “batch analysis of an audio corpus to mark the aural/perceptual relationships of poetry performance and extra-poetic ‘asides’ (which often provide significant contextualization of the poems but which may be ‘invisible’ in the Pennsound archive).”

Another participant interested in the APS recordings wanted to discover what it meant to think through “how a digital archive can recover intangible and ephemeral yet deeply powerful social experiences of sound” including “[w]hat themes of identity, gendered relations, and intercultural relations, may be heard in the Native speakers’ and singers’ expressions and performances of the recorded stories and songs in the collections;” this participant wondered, “how might we thematize and index sounds to address issues of indigenous sonic embodiment in files from which we can hear but not necessarily see the speakers and singers? What are the [sonic] differences and similarities among performers of similar source material? How do these performative differences/similarities map or not map onto other factors (race, gender, region, class, age, etc.)?” Also interested in the APS Native American collections, another participant wanted to analyze these holdings in order to classify Navajo speakers against a map of origin in order to illustrate the location of a speaker. With the ultimate goal of “develop[ing] a cultural map to show spheres of influence of those language-speaking approaches on the stories and motifs across time and in proximity to historical centers of tribal trauma,” this participant wanted to use software “to determine

whether dialectical region or if proximity to historical centers of tribal trauma (e.g. boarding school experiences or Navajo Long Walk) influence that speaker’s . . . Beauty Way and Protection Way approaches to speaking the Diné language”.

Conclusion

Ultimately, can a computer be taught to distinguish between paralinguistic commentaries and formal (or informal) poetry readings or a Beauty Way speaker and a Protection Way speaker within a large collection of sound files? This paper attempts to imagine these possibilities by positioning what users want to do with sound within a critical framework of listening theories that understand “the voice” as a cultural phenomenon that reflects the resonance between linguistic and non-linguistic features of sound. This paper will primarily frame user requirements gathered as part of HiPSTAS within listening theories in the humanities. However, this paper will also briefly mention possible ways forward including a use case in which PennSound poets and scholars use TEI speech tags for tagging tempo, rhythm, loudness, pitch, tension, and voice across PennSound poetry files in order to enable machine learning with ARLO (Adaptive Recognition with Layered Optimization) software, a machine learning application for analyzing sound on Stampede, an NSF petascale HPC system at the Texas Advanced Computing Center. Finally, we need to understand what users want to do with sound and the theories behind critical listening in the humanities before we can design distant listening tools that afford sound scholarship.

References

- Council on Library and Information Resources and the Library of Congress, *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, 2010, 157.
- Jerome McGann** (2008), *Radiant Textuality: Literature After the World Wide Web* (New York: Palgrave).
- Walter J. Ong** (1967), *The Presence of the Word: Some Prolegomena for Cultural and Religious History* (New Haven: Yale University Press), 88.
- Marjorie Perloff and Craig Douglas Dworkin** (2012), *The Sound of Poetry, the Poetry of Sound*, Chicago: University of Chicago Press, 2009; Adalaide Morris, ed., *Sound States: Innovative Poetics and Acoustical Technologies*, Chapel Hill, NC: The University of North Carolina Press, 1998; Jonathan Sterne, ed., *The Sound Studies Reader* (New York: Routledge).
- Roland Barthes** (1978), *Image-Music-Text*. (New York: Hill and Wang), 182.
- Barthes**, *Image*, 182.
- This seems related to **Roland Barthes**’s three distinct types of listening in his essay “Listening”: the first represents a listener on “alert” as prey or predator, as mother or child, as lover on the lookout; the second represents “deciphering” or “what the ear tries to intercept are certain signs”; the third is the “intersubjective” listening of the psychoanalyst. **Roland Barthes**, *The Responsibility of Form*, trans. **Richard Howard** (New York: Hill and Wang, 1985), 245; **Michael Chion**, “Three Listening Modes,” in *The Sound Studies Reader*, ed. Jonathan Sterne, 48-53 (New York: Routledge, 2012), 50 - 51.
- Chion**, “Three Listening Modes,” 51.
- Chion**, “Three Listening Modes,” 50; emphasis added.
- Adriana Caravero** (2012) “Multiple Voices,” in *The Sound Studies Reader*, ed. Jonathan Sterne, 520- 532 (New York: Routledge), 530, 531.
- Mladen Dolar** (2012), “*The Linguistics of the Voice*,” in *The Sound Studies Reader*, ed. Jonathan Sterne, 539-554 (New York: Routledge), 552.
- Dolar**, “Linguistics,” 539.
- Dolar**, “Linguistics ,” 544.
- Please see the project site at <http://blogs.ischool.utexas.edu/hipstas/>.
- Please see the participant list and links to their project interests at <http://blogs.ischool.utexas.edu/hipstas/participants/>.

Project reports are publically available at <https://sites.google.com/site/nehhipstas/project-pages>.

Beyond the Tool : A Reflexive Analysis on Building Things in Digital Humanities

Couture, Stéphane

McGill University, Canada

Sinclair, Stéfan

McGill University, Canada

The goal of this paper is to present some reflections about the process of building things in Digital Humanities. It is based on our own experience in developing an analytic tool to study Internet Relay Chat (*IRC*) conversations within hacker and free and open source software communities.

Questions have been raised recently about the epistemology of building and of built artifacts within Digital Humanities. Following Lev Manovich's provocative statement that a "prototype is a theory", Ramsay and Rockwell have argued that the activity of building a digital prototype should be "capable of providing affordances as rich and provocative as that of writing" (Ramsay and Rockwell 2012, 83). Galey and Ruecker (2010), for their part, propose that the prototype should be received as conveying an argument, as would a book or article, and be evaluated as such. These reflections are interesting in that they propose to go beyond the mere building of a tool to a thinking about things and the building process as valid scholarly contribution. We would like to pursue this line of reasoning but instead of arguing the epistemological validity of tool, we propose to consider the very building process as a methodological and ethnographically-oriented opportunity to reflect on the studied material and the design process. In a sense, we follow Phil Agre's approach, recently re-mobilized by Software Studies theorist Warren Sack (Forthcoming), in pursuing "a technical practice for which critical reflection upon the practice is part of the practice itself" (Agre 1997, xii).

The tool we will present, *IRCMine1*, was developed in the first part of 2013, within a wider context concerned with data mining conversations and interactions in hackers communities (such as free and open source software and Anonymous). Indeed, although many tools were developed (or are still being developed) to study different aspects of free and open source software online communities – tools for the analysis of mailing list, bug trackers or repository commits – the analysis of conversations from IRC logs remains neglected. It is still more important to look at this, since IRC is being used increasingly within free software communities, as an open, synchronous, group conversation protocol. Moreover, IRC is a tool of choice for many hacktivist groups such as Anonymous that coordinate their action in this space (Coleman, 2013).

```
11:15 -!- stephc [-steeph@DigiHum.McGill.CA] has joined #drupal-support
11:15 -!- Irsxi: #drupal-support: Total of 326 nicks [0 ops, 0 halfops, 0 voices, 326 normal]
11:15 -!- Irsxi: Join to #drupal-support was synced in 2 secs
11:15 -!- webflio [-Adium@gateway.digi-info.de] has quit [Quit: Leaving.]
11:16 -!- fatalbert619 [811556ba@gateway/web/freenode/p.129.21.86.186] has joined #drupal-support
11:16 -!- jhedstrom [-jhedstrom@75-150-34-209-Oregon.hfc.comcastbusiness.net] has quit [Quit: Leaving.]
11:16 -!- jhedstrom [-jhedstrom@75-150-34-209-Oregon.hfc.comcastbusiness.net] has joined #drupal-support
11:16 -!- fatalbert619> HEY HEY HEY!
11:17 < fatalbert619> Do anyone have an experience setting up the Google Identity Kit on the site?
11:18 -!- aspilicchio [-aspilicchio@fragger.nascom.be] has left #drupal-support
```

We propose three axes of reflexive exploration about our experience in building the prototype:

1) Reflection about the studied material. The first axis of reflexive analysis concerns the material, and especially the format of the log files. Our design practices brought us to consider more closely the log files format and the form interactions held in IRC channels. For instance, what could be considered as a conversation in IRC files? Considering the close imbrications between metadata and messages (content), do we consider IRC files as a text? Also, the very choice of looking at IRC conversations – instead of mailing lists or commit

logs – can also be reflected upon, since it was justified by the need to look at a less visible space of interactions. In a sense, choosing to give visibility to this space was also a choice about giving visibility to some kind of work over others (Star and Strauss 1999).

2) Ethical aspects of designing the interface. A second set of concerns is related to ethical concerns, such as having a balance between ease of use and keeping the confidentiality of the studied data. Indeed, most of the time, IRC logs are not available publicly and can only be collected by the researcher, thus posing questions about confidentiality of the data. This presented some important conceptual and technical challenges since we decided to develop a web-based tool (JavaScript, HTML, etc.), thus relying on the web browser to execute the code. Although it could be easy for a technical person to install this code on a local machine and ensure the security, how do we design an interface so that users can trust that the data being analyzed will stay confidential? How do we balance usability and performance on one hand, and security on the other? This axis of reflexive thinking is similar to the proposal of a value sensitive design where attention to values and ethical concerns are integrated in the very process of design (Friedman, Kahn, and Bornning 2002; Le Dantec, Poole, and Wyche 2009).

3) Reflection on our design (and coding) practice. This was interesting since one member of our team (Couture) did his thesis on source code, and coding. The project allowed him to experience the actual coding practices (after a long hiatus of coding), especially related to modularity and the circulation of code objects. Although at the start the programming was done in a very ad-hoc manner, it soon became important to modularize the source code and have some consensus on programming standards and the organization of the files. In a way, the organization of source code was articulated to reflect the organization of our collective work. Our coding practice also allowed us to investigate and better understand the different technological resources available to – and often used by – free software coders and other programmers. It would, for instance, allow us to better understand the dynamics around the GitHub Platform, something that has already received some attention by scholars in social science (Takhteyev and Hilts 2010).

This paper will summarize the problematic as well as our objectives in the development of this tool. However, we propose to concentrate most of our presentation on a reflexive analysis of our own design activity in the development of this tool.

References

- Agre, P. E.** (1997). *Computation and Human Experience* (Learning in Doing: Social, Cognitive and Computational Perspectives). Cambridge: Cambridge University Press.
- Coleman, G.** (2013). "Anonymous in Context: The Politics and Power Behind the Mask". Paper No. 3. Internet Governance Papers. http://www.cigionline.org/sites/default/files/no3_7.pdf.
- Friedman, B., P. H. Jr. Kahn, and A. Bornning.** (2002). "Value Sensitive Design: Theory and Methods", UW CSE Tech. Rep. 02-12-01. <http://www.urbansim.org/pub/Research/ResearchPapers/vsd-theory-methods-tr.pdf>.
- Galey, A., and S. Ruecker.** (2010). "How a Prototype Argues." *Literary and Linguistic Computing* 25 (4) (October 27): 405–424. doi:10.1093/lrc/fqq021. <http://llc.oxfordjournals.org/content/early/2010/10/26/llc.fqq021.full.pdf+html>.
- Le Dantec, C. A., E. S Poole, and S. P Wyche.** (2009). "Values as Lived Experience: Evolving Value Sensitive Design in Support of Value Discovery." In Proceedings of the 27th International Conference on Human Factors in Computing Systems, 1141–1150.
- Ramsay, S., and G. Rockwell.** (2012). "Developing Things: Notes Toward an Epistemology of Building in the Digital Humanities." In Debates in the Digital Humanities, edited by Matthew K. Gold. Minneapolis: University Of Minnesota Press.
- Sack, W..** Forthcoming. The Software Arts.
- Star, S. L., and A. Strauss.** (1999). "Layers of Silence, Arenas of Voice: The Ecology of Visible and Invisible

Work."Computer Supported Cooperative Work 8 (1-2): 9–30.
<http://portal.acm.org/citation.cfm?id=309448>.

Takhteyev, Y., and A. Hilts. (2010). *Investigating the Geography of Open Source Software through GitHub*. University of Toronto. <http://takhteyev.org/papers/Takhteyev-Hilts-2010.pdf>.

Validating Computational Stylistics in Literary Interpretation

Craig, Hugh

University of Newcastle - Australia

Eder, Maciej

Pedagogical University of Kraków, Polish Academy of Sciences - Poland

Jannidis, Fotis

University of Würzburg - Germany

Kestemont, Mike

University of Antwerp - Belgium

Rybicki, Jan

Jagiellonian University Kraków - Poland

Schöch, Christof

University of Würzburg - Germany

0. Introduction

Quantitative analysis of literary texts is now well established in authorship attribution. There are continuing lively discussions of method, and the understanding of how classification works best with language continues to evolve, but there are some successes to point to, and most literary scholars accept that when experts disagree on an attribution, a statistical approach can be helpful. The great advantage for quantitative work in this area is that methods can be tested with texts of known origin, so that calibration and validation can be done. Practitioners and traditional scholars can share confidence in rigorous studies with good sampling, controls and validation of various kinds.

There is also the potential for computational stylistics to make a contribution in stylometry beyond authorship attribution and in the wider area of literary interpretation. However, here the problem of validation is acute. If a surprising finding emerges from a quantitative study, how can we tell a chance result, or an artefact of method, from a well-founded finding? How do we judge the robustness of the results and the degree to which conclusions may be generalized? How do we relate analyses based on thousands of long texts to the established understanding of areas of culture based on the intensive study of a few works? The best first options in moving beyond authorship may be other areas of classification, where validation is still possible, like chronology and genre study, but the bigger challenge and greatest rewards will be in interpretation in the wider sense.

This panel approaches the question of validation in computational stylistics beyond authorship attribution through case studies in a variety of languages and literary traditions. Each of the case studies concerns computational stylistics beyond authorship attribution, discusses issues of validation, robustness, and/or interpretation, and offers some considerations of the wider questions of method which arise. The panel will combine brief presentations of the use cases exemplifying the larger issues with ample time for discussion among the panelists and with the audience.

1. Fotis Jannidis & Hugh Craig: Statistical complexity in the language of lowbrow and highbrow novels in German and English

In this study we apply Shannon Entropy and Jensen-Shannon Divergence (Lin, 1991; Rosso et al., 2009) to the language of the novel, using one German and one English corpus from the nineteenth century, and one English corpus from the late twentieth century. We focus on the way low-brow and high-brow novels score on the two measures. We rely on classifications from standard literary histories for the novels and explore the statistical results with close readings of selected passages.

Figure 1 is a graph of Jensen-Shannon Divergence and Shannon Entropy in the more modern English corpus. The novels from the Booker Prize shortlist have generally higher scores on both measures than the other identified groups of low-brow novels (with a p-value of 0.0007 for the t-test on JSD).

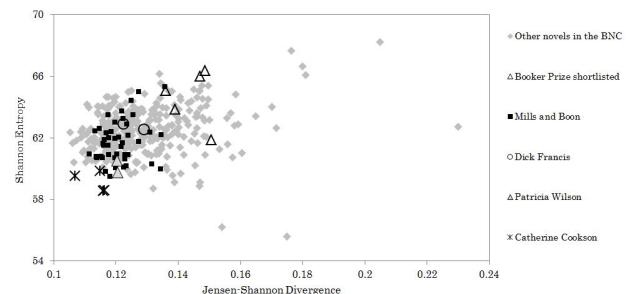


Fig. 1: Jensen-Shannon Divergence and Shannon Entropy scores for 376 novels in the BNC

The pattern for the nineteenth-century corpora is different. In the early- to mid-Victorian English novels, works by Charles Dickens have low scores on both Entropy and Divergence, despite his canonical status. In the 1860-90 German corpus, some popular novels, for example those by Johanna Spyri, author of the famous Heidi books, also have low scores on Entropy and Divergence, but some works by popular authors like Marlitt span the highest to the lowest range of divergence, revealing a hitherto unnoticed variety of styles (the t-test didn't show a significant difference between the groups). Under some circumstances Entropy, which has been judged not to be a useful measure for author-attribution studies (Hoover, 2003), and Jensen-Shannon Divergence seem to be useful to distinguish between lowbrow and highbrow novels, but it is yet unclear under which circumstances. In the paper we will discuss the overall relationship between the information-theory metrics as applied to language and classifications of the novels in terms of market sectors, relying on standard measures of statistical significance to validate our claims.

2. Maciej Eder: Bootstrap consensus network: towards a robust visualization in stylometry

Stylistic methodology, developed to solve authorship problems, can easily be extended and generalized to assess different questions in literary history. Explanatory multidimensional methods, relying on distance measures and supported with visualization techniques, are particularly attractive for this purpose. However, they are very sensitive to the number of features (usually: frequent words) analyzed. Even worse, they are either unable to fit dozens of texts on a single scatterplot (e.g. Multidimensional Scaling), or highly dependent on the choice of a linkage algorithm (e.g. Cluster Analysis). The technique introduced in this study combines the concept of network as a way to map large-scale literary similarities (Jockers, 2013), the concept of consensus (Lancichinetti and Fortunato, 2012), and the assumption that textual relations usually go beyond mere nearest neighborship.

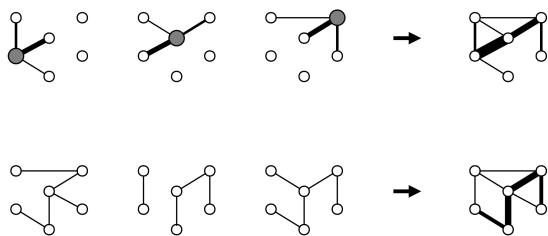


Fig. 2: Two algorithms of mapping textual relations

Particular texts can be represented as nodes of a network, and their explicit relations as links between these nodes. The procedure of linking is twofold. One of the involved algorithms (Fig. 2, top) computes the distances between analyzed texts, and establishes, for every single node, a strong connection to its nearest neighbor (i.e. the most similar text), and two weaker connections to the 1st and the 2nd runner-up (i.e. two texts that get ranked immediately after the nearest neighbor). The second algorithm (Fig. 2, bottom) performs a large number of tests for similarity with different number of features to be analyzed (e.g. 100, 200, 300, ..., 1,000 MFWs). Finally, all the connections produced in particular "snapshots" are added, resulting in a consensus network. Weights of these final connections tend to differ significantly: the strongest ones mean robust nearest neighbors, while weak links stand for secondary and/or accidental similarities. Validation of the results – or rather self-validation – is provided by the fact that consensus of many single approaches to the same corpus sanitizes robust textual similarities and filters out apparent clusterings. The idea discussed in this paper can be applied to map large literary corpora (see the contribution by Jan Rybicki, below).

3. Jan Rybicki: Validating a large bootstrap consensus network in literary history

Over five hundred English novels from Swift to Rowling were used to produce a bootstrap consensus network of most-frequent-word frequencies, using a pseudo-bootstrapped cluster analysis from 100 to 1,000 most frequent words with the stylo package (Eder et al., 2013) for R and visualized with GEPHI's Force Atlas 2 layout algorithm (Bastian et al., 2009). The resulting graph yielded the usual strong authorship signal, but the overall shape exhibited a number of features that make sense in the context of traditional literary history.

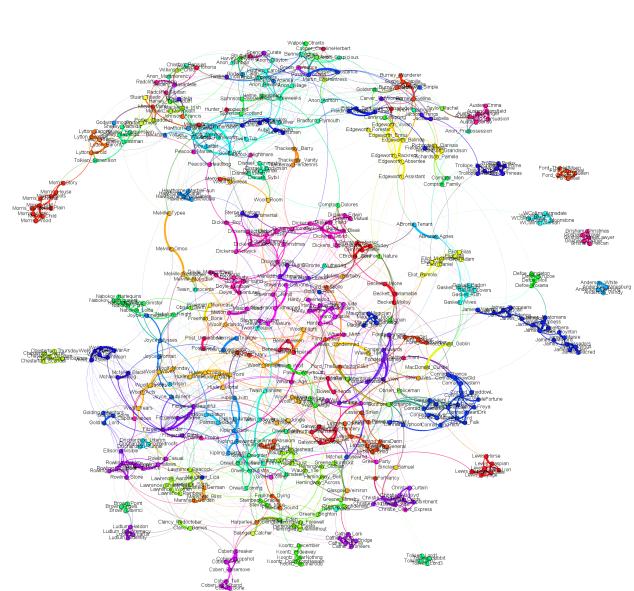


Fig. 3: Bootstrap consensus network of over 500 English novels

The overall shape of the resulting network (Fig. 3) is reminiscent of the earlier-observed "stylistic drift" phenomenon (Burrows, 1994) in its general chronological order and observes certain topographic rules. The texts are roughly ordered from top (early) to bottom (late), with three avenues of transition from the top-most 18th-century writings to late Victorians and modernists: Americans Melville and Hawthorne; Dickens; and the mid-Victorian female writers, with some notable outliers. Most of the latest texts in this set gravitate towards the bottom area of the graph.

With the impact of spelling variation minimized by 100% culling (a procedure by which all words that do not appear in each of the studied texts are rejected from the analysis), this is a clear indication that distant reading by most-frequent-words frequencies can mirror the evolution of literary style over hundreds of texts and hundreds of years and open new perspectives for close reading. After all, the application of statistics to literature is remarkable in that its results can be validated not by statistical means alone, but also, and perhaps above all, by traditional literary history, classification and interpretation.

4. Christof Schöch: Validating and interpreting Principal Component Analysis: A Case-Study from the Analysis of French Enlightenment Plays

This case study investigates issues of validation and interpretation of stylometric results with regard to authorship, genre, date and form, based on a collection of 120 French plays from the French Enlightenment period. Preliminary analyses using Cluster Analysis have suggested that besides authorship, categories like genre (tragedy or comedy) and form (verse or prose) are important stylistic signals in this collection.

To verify this observation, PCA (Jackson, 2003; Diana & Tommasi, 2002) was performed on different subsets of plays. In a subset of 19 comedies in either verse or prose written by five authors between 1712 and 1760, PCA shows strong effects for form (verse or prose) but results appear to vary for different settings (particularly, number of frequent words).

To test the salience of the effects for form and the robustness of the results, the contribution of several variables to the first three principal components was calculated for different settings. More precisely, F-measures of ANOVA tests were calculated for author, form and date in relation to PC1 to PC3 for PCAs based on 5-200 most frequent words.

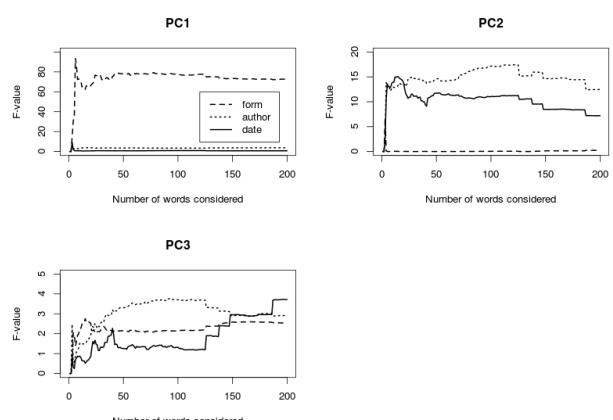


Fig. 4: F-scores for author, form and data on PC1 to PC3

Figure 4 shows how PC1 is dominated by "form", with an extremely high F-score, while "author" and "date" hardly contribute. In PC2, form does not play any role, but "author" and also "date" do. Considering its reduced scale, it appears that PC3 does not show any clear trends.

That the effect for "form" is so concentrated in one major component comes as a surprise, especially because similar effects have not been observed in other domains, such as Early Modern English Drama (Hugh Craig, paper in preparation). However, given the striking robustness of the results, the

interpretation of the PCA can proceed with confidence. Further application domains of this method are stylometric investigations of variables such as genre, theme or literary period.

5. Mike Kestemont: Learning Deep Representations of Characters in Literary History

One of the most exciting movements in current Machine Learning is “Deep Learning” (Bengio, 2009). In this field, people attempt to leave the idea of “hand-crafted” features. Older, so-called “shallow” learning techniques – commonly used in stylometry – heavily depend on a researcher’s, typically strongly biased, representation of a problem and will not attempt to optimize or even correct this representation. In “Deep Learning”, the idea is that one should not only learn how to solve a problem, given some input information, but additionally, how the input information is best represented in order to solve the problem. To achieve this, researchers send data through a layered structure of units (“neural network”). At each subsequent layer in this network architecture, the representation of the original input information grows increasingly abstract.

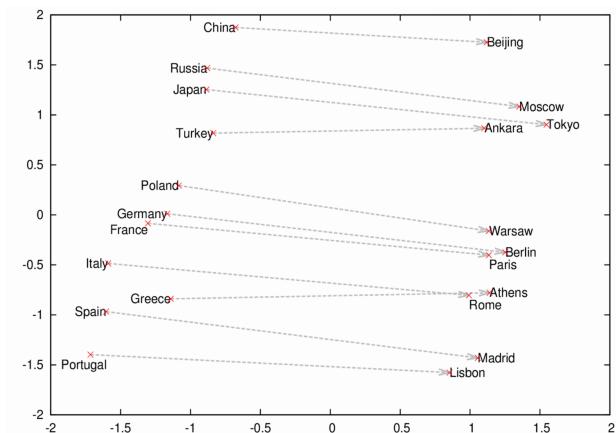


Fig. 5: Country and Capital Vectors Projected by PCA. (Copyright: Mikolov et al. for Google Inc.)

Recent research has demonstrated that Deep Learning yields extremely valuable problem representations. In distributional semantics, for instance, it yields an extremely powerful vector-space model of words. This contribution will survey how such vector spaces have recently been used for advanced analogical reasoning. In a series of breakthrough papers, Mikolov et al. (e.g. 2013) have shown that these models (see Fig. 5) can be used to answer complex questions like: “What is to king, like woman is to man?” (answer: “queen”), or “What is to Warsaw, like France is to Paris?” (answer: “Poland”). I will discuss how this kind of representational learning could be applied to modeling characters from literary history. The main idea is that we should be able to easily answer questions about the archetypical relationships between characters: Who is to Romeo, like Isolde is to Tristan?”. I will argue that this approach offers an exciting new framework to study the “meaning” of literary personas (cf. Bamman et al. 2013) and their cross-novel interrelations.

References

- Bamman, D., Brendan O'Connor, B. and Smith, N.** (2013). *Learning Latent Personas of Film Characters*. ACL 2013, Sofia, Bulgaria. 352-361.
- 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.
- Bengio, Y.** (2009). *Learning Deep Architectures for AI*. Foundations and Trends in Information Retrieval 2: 1-127.
- Burrows, J. F.** (1994). *Tiptoeing into the infinite: testing for evidence of national differences in the language of English narrative*. Research in Humanities Computing, 2: 1-33.
- Diana, G., and Tommasi, Ch.** (2002). *Cross-validation Methods in Principal Component Analysis: A Comparison*. Statistical Methods and Applications 11/1: 71-82.
- Eder, M., Kestemont, M. and Rybicki, J.** (2013). *Stylometry with R: a suite of tools*. Digital Humanities 2013: Conference Abstracts. Lincoln: University of Nebraska-Lincoln, pp. 487-89.
- Hoover, D.** (2003). *Another perspective on vocabulary richness*. Computers and the Humanities, 37: 151-78.
- Jackson, J.E.** (2003). *A User's Guide to Principal Components*. Hoboken: Wiley.
- Jockers, M.** (2013). *Macroanalysis: Digital Methods and Literary History*. Champaign: University of Illinois Press.
- Lancichinetti, A. and Fortunato, S.** (2012). *Consensus clustering in complex networks*. Scientific Reports, 2: 336: 1-7.
- Lin, J.** (1991). *Divergence measures based on the Shannon entropy*. IEEE Transactions in Information Theory, 37(1): 145-51.
- Mikolov, T., Yih W. and Zweig, G.** (2013). *Linguistic Regularities in Continuous Space Word Representations*. Proceedings of NAACL-HLT 2013, Atlanta, Georgia, 746-751.
- Rosso, O., Craig, H., and Moscato, P.** (2009). *Shakespeare and other English Renaissance authors as characterized by Information Theory complexity quantifiers*. Physica A, 388: 916-26.

Readings of a photograph: Cognition and Access

Das Gupta, Vinayak

Trinity College Dublin, Ireland

In the act of interpreting and describing pictures, even in the fundamental process of cognition, there is a strong play of language in the visual field. Critics working in this field have described the same relationship between the word and the image through different phrases: what Foucault refers to as the “seeable” and the “sayable” (1982) is defined by Deleuze as the “display” and the “discourse” (1984) while W.J.T. Mitchell terms as the “showing” and the “telling” (1994). The study of word and image (painting and poetry, literature and the visual arts), their relationship, or the examination of culture using critical devices in each field has been a consistent theme in the literary fields since antiquity. Visuality requires verbal descriptives for interpretation, whether they are spelt out explicitly, or subconsciously attributed in the human mind.

Images, when committed to the digital space, pose new issues regarding cognition. How do we recognise the digital object outside the moment of experiencing it? The digital collection is an assimilation of filenames and to use the digital object, we must first be able to recognise it. An image file may be recognised through the filename extensions (JPEG, TIFF etc.), but to conclusively state that it is the digital image of a photograph cannot be done without first looking at the contents of the digital file. The digital object does not have a tangible form or lineament. Physical photographic artifacts reveal clues to determine different aspects of its source. The different material on which the photographic image is imprinted (glass or paper) can reveal clues towards the origins of the image. Since photography is as much a technological phenomenon as it is fruit of human endeavour, the physical object itself communicates moments of technological change. In the digital medium, however, the photo-ancestry is lost and a new inscription formed. How do we then, perceive the photographic artifact in its computerised form? The question of recognition is central to the argument of reading the digitised image.

How do we, as spectators of the photograph, read the image? The photographic image bears a likeness to its subject (icon) and is a physical extension of it (index). The photograph possesses an evidential force: it cannot be argued that what it captures, through the process of light falling on a photo-

sensitive plate, was not there. The readings of a photograph is dependent on the layers of recognition that happen in the process of viewing the image. The recognition of the indexical contiguity is directly related to the spectator's familiarity of the photographic subject. In the second instance, all images demand a recognition of purpose: this purpose can be the photographer's own, or it could be one that the photograph creates for itself -- a new life for its subject. The photograph can only depict history in a bounded frame: it is unable to speak. Thus the recognition of purpose of the image provides readings into the contextual framework within which the image is placed. *Slaughter Ghat, Cawnpore* (BL, Photo 193/20) presents us with a topographical space -- a river bank where several small shrines border upon a lower mud shelf above the river, on the banks of which are tethered two country boats. The intention of the photograph is not to portray a simple country scene (though it might) but to draw our attention to the site of a brutal massacre where hundreds of British refugees were fired upon and killed by rebelling forces in India. The purpose, then, becomes an instrument in the reading of the photograph. Reinforcing this view, John Berger (1972: 10) writes:

"Every image embodies a way of seeing. Even a photograph. For photographs are not, as is often assumed, a mechanical record. Every time we look at a photograph, we are aware, however slightly, of the photographer selecting that sight from the infinity of other possible sights."

The third and final instance is that of a recognition of source and this is of significant importance for the archival image. The little that we can articulate about the photographic image is derived from an understanding of the history of the object (the physical photograph). How we proceed to classify the image, place it within numerous other photographs is dependent on how we recognise the photographer, the period, the photographic plate and the photographic process. The history of the object is vital in our attempts to place it within the structures of a digital collection.

The naked digital artifact is wrapped in an envelope of tags and markers in an attempt to locate and describe the object. The categories described within the catalogue are translated in the digital medium as metadata. Metadata standards establish a common understanding of meaning or semantics of data. This aids proper use and interpretation of the data by its owners and users. My paper will explore the formation of metadata through the analysis (a combination of statistical and close-reading) of a body of annotated photographs. I propose to demonstrate means of extracting meaningful metadata which addresses the theoretical issues stated above, in order to place them within established, standardised formats. I will also demonstrate new methods of visualising large image collection through the use of this form of analysis. While my paper will focus on early photography from colonial India (1850-1900), the scalability of such a project will also be a point of consideration in this paper.

References

- Berger, John (1971). *Ways of Seeing*. London: British Broadcasting Corp. Print.
- Deleuze, Gilles (1984). *Kant's Critical Philosophy: The Doctrine of the Faculties*. University of Minnesota Press
- Foucault, Michel (1982). 'The Subject and Power', *Critical Inquiry*, Vol. 8, No. 4 (Summer, 1982). London: The University of Chicago Press, pp. 777-795
- Gitelman, Lisa (2006). *Always Already New: Media, History, and the Data of Culture*. Cambridge, MA: MIT Press.
- Mitchell, W.J.T. (1994) *Picture Theory*. London: The University of Chicago Press
- Peirce, Charles S. (1940) 'Logic as Semiotic: The Theory of Signs', *Philosophical Writings of Peirce*. Edited by Justus Buchler (New York: Dover, 1955)
- The issue of committing to the digital sphere brings to mind the question of virtual inscription. While we are acutely aware of the way the photographic image is inscribed, digital inscription moves in more mysterious ways. Lisa Gitelman writes: "I have tended to chalk this up to the difference between the virtual and the real, without stopping to ponder what virtual inscriptions ... could possibly be. Like the mysteries surrounding

the inscription of recorded sound onto surfaces of tinfoil and then wax at the end of the nineteenth century, the mysteries surrounding the virtual inscription of digital documents are part of the ongoing definition of these new media in and as they relate to history." (2006: 19)

The term index (devised by the American philosopher Charles S. Peirce) is often used to describe this quality of causal contiguity. Peirce distinguished indexes from icons and symbols, writing that the index 'refers to the Object that it denotes by virtue of being really affected by that Object' For more see, Peirce (1955 :102)

The identification of purpose lies outside the photographic frames, through there may be visual clues within the image. The purpose determines the context.

Amongst collectors, identifying the first (existing) print from the negative is of importance, for value as an object of collection lies, ironically, in the 'uniqueness' of the first print. With digital reproductions, it is more difficult to ascertain the lineage of the original photograph.

The Scholarly 3D Toolkit: Annotation, Publication, and Analysis of 3D Scenes alongside Imported Humanities Data

DataColtrain, James Joel

University of Nebraska, United States of America

New advances in online game engines have made it possible to easily view 3D virtual environments from any web browser, but the full potential of 3D humanities research has gone unrealized because of the difficulty in connecting important 3D findings to the work of traditional scholars grounded in texts. This presentation will discuss the current development and show demonstrations of the Scholarly 3D Toolkit, (S3DT) a plug-in for the Unity game engine designed to help better interface 3D historical reconstructions with other data. The work of a team lead by James Coltrain, S3DT will provide simple interfaces that allow creators to link their 3D scenes to sources and documents, and to dynamically import and view traditionally indexed digital humanities data from databases, spreadsheets, or GIS programs. The result will allow users to view multiple layers of data plotted within a single online 3D environment, showing markers for events, personal connections, documents, images, and annotations from multiple users, all in time and space. S3DT will allow for greater and more sophisticated interdisciplinary analysis, helping scholars studying three dimensional spaces to contextualize models of architecture, urban structures, and natural topography using texts and other spatial data. By comparing existing digital humanities findings with 3D scenes that show scale, light, and texture, the platform will allow for more complex and nuanced investigations of past spaces. Along with a discussion of the project's progress and the theoretical questions at play, this presentation will show early demos of a test case for the platform. These will include a richly annotated high quality 3D reconstruction of Fort Stanwix, an 18th-century historic site and National Monument, with an existing database constructed by Nebraska undergraduates of over 400 letters, maps, and plans.

S3DT will build upon the achievements of previous digital humanities projects by expanding the options scholars have for working in 3D spaces. Earlier platforms have allowed for the real-time display of annotated 3D models, but some could not stream live in a browser, and most allowed creators little in the way of customization.ⁱ Extremely important work has been done with diverse and creative applications of historical GIS, and S3DT will allow for those established types of analyses to be brought into the third dimension. ⁱⁱ More recently, some scholars have made use of online game engines like Unity to achieve some of the goals set forth in the S3DT project, including the use of advanced real time graphics in an online environment. ⁱⁱⁱ However, these projects have not resulted

in open, customizable platforms, and none allow for the importation of new 3D content. S3DT will build upon previous work in Unity by connecting 3D scenes from multiple creators to the layered viewing of all kinds of outside humanities data.

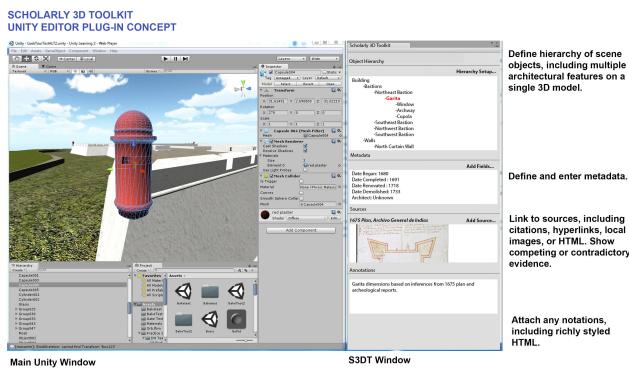
The practical tools in S3DT also make many previously difficult modes of spatial analysis quicker and more accessible. S3DT scenes can show spaces changing over time with numerous iterations and nuance, and also display multiple interpretations of the same structure side by side as competing arguments. With 3D objects linking to multimedia sources, users can now better understand the interpretive leaps creators made, and which pieces of fragmentary evidence scholars privileged in creating coherent 3D spaces, information that also facilitates efficient peer review. The ability to display different types of data can also promote public outreach in addition to academic collaboration, letting universities, museums, archives, historic sites, and even individual visitors contribute to the same online 3D spaces. The design of the S3DT plug-in for Unity will allow for open analysis of 3D scenes, while protecting scholars' data for future use. The plug-in does not interfere with the traditional workflows for 3D content creation, and also stores all textual and multimedia data in standard MySQL databases. As a result, neither scholars' 3D models nor their annotations or data will become stuck in the S3DT if creators find better future platforms for presentation.

S3DT will consist of a two part plug-in for Unity. The first part, within the Unity Editor, will allow creators to add notes and metadata to imported 3D objects, and prepare them for publishing. The second, is a web template which will allow for the viewing and manipulation of published scenes, as well as the live importation and plotting of new data layers from outside sources. Below is a typical workflow for S3DT along with key features at each step. This presentation will conclude with early demonstrations of many features from both parts of the plug-in.

I. 3D Content Creation

Creators begin by modeling and texturing a 3D scene in their typical workflow in a 3D suite such as 3D Studio Max, Maya, or Blender. When they have finished, they export their 3D models to industry standard formats (ex. .obj or .fbx). They then download and install the Unity Editor and the accompanying S3DT plug-in. Finally they load their 3D scene into the Unity Editor.

II. S3DT Plug-in for Unity Editor



With their models loaded into the Unity editor, creators will use the S3DT plug-in to prepare them for publication. This includes creating an object hierarchy, denoting nested neighborhoods, complexes, buildings, rooms, architectural features, and sub-features, each as defined by the creator. Once the objects are defined, creators can enter metadata for each scene object in any fields they like. In particular, creators will be able to enter time sensitive information, such as the dates for the object's creation, alteration, damage, and destruction. Creators will also be able to enter links to sources used in their interpretation of the reconstructed object, as well as notes about their specific decisions.

Next creators will publish their scene. In the process each published scene is exported as two parts, a Unity 3D file formatted for web display, and a matching MySQL database containing all metadata and links to sources and annotations.

III. S3DT Plug-in for Browsers



The S3DT web browser plug-in consists of a Javascript library and web template for loading and displaying published S3DT scenes on the web. Most projects will use a customized version of the web template, but the Javascript library is available for projects that are integrated into existing sites or for which creators desire a higher level of customization.

The S3DT browser plug-in has a simple user interface consists of the following:

- The **Main Window** displays the published Unity scene in real-time 3D.
- The **Timeline** consists of a scalable time line with a slider to control time position and markers corresponding to time sensitive events plotted in the scene.
- The **Layers List** shows all the elements in the scene, organized by package. Each layer has a collapsible view that expands to show the entire object hierarchy as defined by scene creators in the S3DT Unity Editor plug-in. Any additional content or data loaded into the 3D scene will appear in the layers list as a new layer, including published S3DT packages, maps, images, collections of user annotations, collections of plotted events, GIS data, etc. Users will be able to toggle the visibility and opacity of any layer or any object within a layer hierarchy.
- The **Tools Window** features a set of utilities users can use to manipulate or analyze the scene. These will include:
 - **Advanced Search** - Allowing customizable complex searches bases on any metadata field or object attribute.
 - **Groups** - Allows users to group objects from any layer together into a new layer.
 - **Edit Object Metadata** - If enabled, allows users to add to or edit metadata for scene objects. **Annotate** - Allows users to leave comments live in the scene, either attached to scene objects or in freestanding 3D space.
 - **Camera Tools** - Allows users to place custom cameras, define camera paths and animate them, and to take and save camera snapshots.
 - **Import Data** - Allows users to import data with geographic information from outside sources including SQL, Excel, KML, and ARCGIS. Also allows users to choose and customize marker appearances based on imported data, or upload their own.
 - **Create Exhibits** - A set of sub-tools will allow users to connect camera views, and animations over time and space with HTML text for guided tours and other exhibits. Created exhibits will load into the layers view.
 - **Map and Image Import** - Allows users to import maps and images into the live scene, and to align maps to existing terrain or images to camera views. Imported maps and images can then load into the layers view.

References

VSIM <https://idre.ucla.edu/gis-visualization/vsim>; Rome Reborn <http://romereborn.frischerconsulting.com>; CDI Second Life projects <http://wt-dc19-prod.astate.edu/a/centers-programs-and-institutes/cdi/projects.dot>.

Anne Kelly Knowles *Past time, past place: GIS for history; David J. Bodenhamer, John Corrigan, and Trevor M. Harris* *The Spatial Humanities*; Spatial History Project <http://www.stanford.edu/group/spatialhistory/cgi-bin/site/index.php>; Hypercities <http://hypercities.com/>; Neatline <http://neatline.org/about/>; World Map Project <http://worldmap.harvard.edu/>. lab, UCLA Experimental Technologies Center <http://etc.ucla.edu/research/projects/romelab/>; Hadrian's Villa Simulation <http://idialab.org/nsf-funded-virtual-simulation-of-hadrians-villa/>; Digital Pompeii <http://classics.uark.edu/DigitalPompeii.html>; Simulated Environment for Theater, <http://humviz.org/set/index.html>.

Digital Cultural Heritage and the Healing of a Nation: Digital Sudan

Deegan, Marilyn

marilyn.deegan@kcl.ac.uk
King's College London

1. Introduction

Sudan is one of the most diverse and culturally rich countries in the world. It is ethnically diverse: the Sudanese are divided among 19 major ethnic groups and about 597 subgroups and speak more than 100 languages and dialects. It is also culturally diverse: tradition, ceremony, language, poetry, art, drama, music and dance, are all vital cultural practices, and Sudan is one of the richest countries in Africa in archaeological remains. Sudan's cultural riches rival those of Egypt, Greece and Rome, but war, famine, displacement and the ravages of time, climate and lack of funds means that the cultural heritage of the country is under severe threat. The preservation and recovery of cultural heritage through digitisation is well-understood by the Sudanese, and many outstanding projects exist throughout the world for Sudan to draw upon. The world knows much about other ancient civilisations, but not much about Sudan. Digitisation will help show the riches of Sudan to the world--and to itself. Many citizens are ignorant of the greatness of the history of their country, and schoolchildren and their elders can benefit greatly from access online to their rich heritage.

The digitisation of selected material of cultural heritage is a national initiative led by the Sudanese Association for the Archiving of Knowledge (SUDAAK), a Sudan-based NGO, to guarantee the long-term preservation, integration, authenticity and accessibility of important cultural content in respective concerned national institutions. The project addresses some of the main issues related to digitisation networks and services in the cultural domain. It specifically aims at safeguarding and reinforcing Sudanese cultural heritage through new technologies. In its initial stage the project will aim at identifying and facilitating the urgent needs for the implementation of appropriate applications of digital technology in cultural content storage and sustainability.

SUDAAK is a cultural non-governmental organisation (NGO) concerned with archiving Sudanese life in history, politics, folklore and culture. While the term archiving is mostly associated with records, the role envisioned for SUDAAK is organising the discovery and display, the celebration and preservation of the traditional and modern knowledge together with the achievements attributable to imagination and leadership of those who were pioneers in laying the foundations of the Sudan and its political, social, economical and cultural strengths. Their major programme now is the Archiving of the

20th Century Sudan Intellectual Heritage, but all other periods and all types of artefacts are within SUDAAK's scope.

1.1. Overview

Digital Sudan

The overall goals of Digital Sudan are:

- Storage of selected recorded cultural material on Sudan within a well-designed selection policy;
- Electronic treatment of old and decaying books and pictures;
- The facilitation of accession to Sudan folklore related material reserved in prominent research institutions;
- Facilitation of access to National Library content needed for Government processes and decision-making;
- Improvement and enhancement of digitisation facilities and services in Sudan;
- The creation of an online national library serves as a model for integrating multi-format and multi-lingual resources from museums, archives, libraries, and bibliographic and Web resources and develops retrieval capabilities;
- Development of a collaborative infrastructure that can support an increasing number of contributing partners nationally; and
- User provision of integrated digital materials that seamlessly link all types of resources.

The key stakeholders are currently:

- National Record Office / Ministry of Council of Ministers
- University of Khartoum/ Ministry of Higher Education & Scientific Research
- Sudan Radio Corporation/ Ministry of Culture and Information
- Sudan National Television/ Ministry of Culture and Information
- National Corporation for Archaeology & Museum/ Ministry of Tourism
- Photography Unit/ Ministry of Culture and Information
- Film Production Unit/ Ministry of Culture and Information
- National Library of Sudan/ Ministry of Culture and Information
- National Research Centre - Information& Documentation/ Ministry of Science and Communication
- Sudan Folklife Documentation Centre / Ministry of Culture and Information
- Africa City of Technology/ /Ministry of Science and Communication
- Sudanese Association for Archiving Knowledge: / Non for Profit Civil Society Organisation

SUDAAK is also working with institutions outside Sudan with expertise in digitisation and digital library development. Currently these include Durham University, with whom SUDAAK have a Memorandum of Understanding, and the Department of Digital Humanities at King's College London, where there is a great deal of expertise in all aspects of this area. In April 2013, the stakeholders listed above were formally constituted as the National Cultural Heritage Digitisation Team (NCHDT). SUDAAK is also planning to work with other institutions world-wide in the development of the plans for the Digital Library.

The content available for digitisation is rich and diverse: In the National Archives alone, there are 76 million photographic negatives recording all aspects of life in the Sudan over the past 100 years. The university library has priceless manuscripts from the beginning of Islam; there are 9 museums throughout the country with artefacts from more than 4000 years of history; film, radio tapes and video record all the major events in the country, as well as the music, dances and traditional practices. Traditional foods and medicine are of great importance too, and there are samples, photographs and documents concerning these in the archives.

Sudan has a good education system overall, with a high level of participation in urban areas. Literacy rates are relatively high, though both participation and literacy rates are lower outside urban areas. The universities are excellent, and there is a modern Open University, established in 2003, that has links to the UK's Open University and the University of Cambridge. The Open University uses all forms of modern

technology to communication with students: video conferencing, Skype, Facebook, websites, as well as radio, television and telephones.

In planning for Digital Sudan, the country has both advantages and challenges. In terms of advantages, the country has an excellent tele-communications infrastructure. It is modern, well-designed, robust and spacious. Sudatel, the main tele-communications company and the National Information Center can provide some of the storage, connectivity, and band-width that should be needed for Digital Sudan, and as the resource grows, the capacity can be increased. In the Ministry of Information and the cultural institutions there is already some technical knowledge, and more importantly, there is huge enthusiasm for the project and a willingness to make things happen. The National Library, National Archives, and the National Museums have good catalogues in place: these are the backbone of any digital library. There are a number of digitisation projects already being undertaken in the cultural institutions and the universities: for example, the University of Khartoum holds the Electronic Sudan Library which provides rare Sudanese materials of historic and cultural significance, with full text that can be searched in Arabic, English and other languages. Sudan Radio has already digitized 27,000 hours of historic radio tapes; the National Museum has digital images of artefacts attached to a catalogue records. But there is much to do, and many challenges and risks. Digital preservation, for example, needs serious consideration. Here, though, Sudan can benefit from excellent work being done in this area by major institutions throughout the world: the US Library of Congress; Europeana; the British Library; the National Library of Wales and many other institutions.

The condition of the analogue materials is also a serious consideration. An intense programme of physical conservation is needed alongside any digitisation activities, and storage of the valuable original artefacts in better conditions than at present is an urgent need.

Next steps

SUDAAK and the NCHDT, together with their international partners, are in discussion with the Ministry of Information and with other funders to identify possible sources of funding for the activity. They are also taking some steps towards training staff in digitisation skills, and digital library development. A major new development is the signing of an agreement with the University of Bergen, Norway, to digitise the archive radio, TV and film materials of the Sudan Radio and Television Corporation.

Conclusions

For a country to embark upon a program as ambitious as this is a huge challenge, and will be costly. Even more costly would be the risk of doing nothing. Sudan is emerging from strife and division into the modern world, and is moulding its new identity by building on the strengths of its cultural memory. Digital Sudan has a huge role to play in this.

References

- Ishma'il Kushkush** (2013), *Ancient kingdoms in land of war*, Khartoum Journal, Africa: New York Times, 31 March 2013, www.nytimes.com/2013/04/01/world/africa/in-sudan-archaeologists-unearth-ancient-kingdoms.html?pagewanted=all
- Simon Tanner** -(2013), *African Manuscripts - a treasure in danger?*, When the Data hits the Fan! The blog of Simon Tanner, Monday, 28 January 2013 simon-tanner.blogspot.fr/2013/01/african-manuscripts-treasure-in-danger.html .
- Marilyn Deegan** (2013), *Digital Sudan: cultural heritage revived & preserved*, When the Data hits the Fan! The blog of Simon Tanner, Wednesday, 24 July 2013, <http://simon-tanner.blogspot.fr/2013/07/digital-sudan-cultural-heritage-revived.html>

tanner.blogspot.fr/2013/07/digital-sudan-cultural-heritage-revived.html

Mapping and Unmapping Joyce: Geoparsing Wandering Rocks

Derven, Caleb

caleb.derven@ul.ie

University of Limerick

Teehan, Aja

aja.teehan@nuim.ie

An Foras Feasa, National University of Ireland, Maynooth

Keating, John

john.keating@nuim.ie

Dept. of Computer Science, NUI Maynooth, Aja Teehan, Cognitive Corp

1. Introduction

The copyright expiry on James Joyce's *Ulysses* in 2012 created a unique opportunity to read the seminal modernist text through the refraction of technologies made available by the Digital Humanities and techniques from Computer Science. *Ulysses* is avowedly and manifestly a work both constructed by and read through explicit references to geography and spatial relations. For instance, Frank Budgen attributes the following statement to Joyce, "I want," said Joyce, as we were walking down the Universitätstrasse, 'to give a picture of Dublin so complete that if the city one day suddenly disappeared from the earth it could be reconstructed out of my book.'"¹

However, it has been suggested that uncertainty and disorientation play as great a part as explicit references to place and these qualities are evoked through specific narrative strategies.² From a Digital Humanities perspective, being able to note such contested or defamiliarising areas presents a challenge.

Significant work has been done in the scholarly literature to manually compile and list named entities such as geographic and place name references in *Ulysses*. However a significant occasion exists to exploit techniques such as XML mark-up and Natural Language Processing (NLP) to explicitly render geographic and spatial references in *Ulysses*, make the references available for machine processing and accessible to users for reading the novel.

This paper investigates the automatic extraction of toponyms from the *Wandering Rocks* episode of *Ulysses*, proposes a model for encoding the episode and accounting for different types of place (including uncertain locations) and, combining these elements, explores XSLTs and visualisations that support a spatial reading of the text. The model proposed by the paper supports not only the notion of the significance of place but also qualities of spatial uncertainty and disorientation noted in the critical literature. The approach taken in the paper leverages existing models and technologies.

2. Methodology

This approach seeks to instantiate geographical evidence in the narrative that is almost exclusively transmitted to the reader through unstructured text in print presentations of the novel. This has been done through a combination of Natural Language Processing tools, geocoding the resulting data and merging the data into a TEI encoded version of the text and presenting the output in a web application.

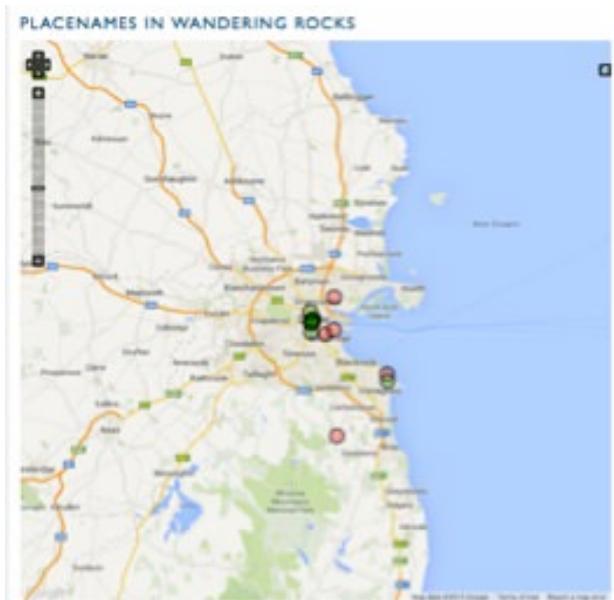
From a Digital Humanities and Computer Science perspective, a number of readily available tools, technologies and methodologies exist to link place names in unstructured text to geographical data. Such tools allow the novel approaches suggested by Moretti and undertaken by Clement.³

¹For example, Named Entity Recognition (NER), a subset of

Natural Language Processing, represents a viable methodology to extract toponyms from unstructured text. Geoparsing place names to match geographical coordinates is assisted through such openly available digital gazetteers such as GeoNames.⁵ Projects such as the University of Edinburgh's Unlock provide non-technical interfaces that allow for automated NER and geoparsing.⁶

232 rows						
Show as: rows		records	Show: 5	10	25	50
All	entityId	type	placeName	settlement	country	geo
1.	wrl_1	route	Prestonby Steps	Dublin	Ireland	
2.	wrl_2	projectedSpace	Ahane	Dublin	Ireland	53.383008, -6.205133099999999
3.	wrl_3	route	Convent Of The Sisters Of Charity	Dublin	Ireland	53.35705049999999, -6.2597354
4.	wrl_4	route	Mountry Square	Dublin	Ireland	53.366981, -6.2175293
5.	wrl_5	projectedSpace	Boston	Derbyshire	United Kingdom	53.23001, 1.9110
6.	wrl_6	projectedSpace	Bvedere	Dublin	Ireland	53.366981, -6.1234688
7.	wrl_7	projectedSpace	London	Dublin	Ireland	53.340143, -6.247436999999999
8.	wrl_8	setting	Ireland	Dublin	Ireland	53.349013, -6.2163097
9.	wrl_9	route	Mountry Square	Dublin	Ireland	53.366981, -6.2175293
10.	wrl_10	marker	Bvedere	Dublin	Ireland	53.366981, -6.1234688

This paper addresses the role that names play in Ulysses, specifically the Wandering Rocks episode and what this role reveals about the novel as a whole. It confronts whether there is a topographical quality in Ulysses and if so how that quality is defined. Accordingly, it considers three hypotheses: that geoparsing of Ulysses enables distant reading that will in turn enable new interpretations of the text; that geoparsing Ulysses creates a virtual gazetteer of the text; that the development of a model for encoding encompasses areas, such as uncertainty as a quality of place, outside of the scope of geoparsing.



Geoparsing and geocoding have been utilised as a primary methodology for the project. A number of technologies such as the Natural Language Toolkit and software produced by Stanford's Natural Language Processing Lab were available and assisted in determining whether such an approach was feasible⁷ ⁸. Again, it was anticipated that an iterative approach would be followed where initial automated extraction of toponyms via NLP would inform the encoding of episodes. This encoding, in turn, would provide the basis for a coterminous presentation of text and geographical elements through the web application and as the basis for interrogating the text along a geographical orientation.



The data model for the Literary Atlas of Europe was utilised as a framework to assign types to toponyms in the resulting TEI.⁹ The framework provided by the LAE allowed for further analysis on the role of place in the text, particularly the representation of uncertainty. Accordingly, while the LAE model provides a preliminary scaffolding, the mode proposed in the paper combines semi-automated extraction of toponyms combined with a document-based encoding.

3. Lessons Learned

The work described by the paper resulted in a number of outcomes. Firstly, the development of the model and subsequent encoding of toponyms in the text rendered a comprehensive and programmatically presentable list of geographic references. Such a list constitutes a sort of virtual gazetteer for the novel. Secondly, this approach works towards identifying any inconsistencies in Joyce's use of geographic references (with regard to the "traps" identified by Hart), indicates the role of geographical uncertainty in the episode and potentially suggests productive interpretive approaches.¹⁰ Thirdly, such an approach contributes towards the notion of a literary cartography and echoes the work undertaken by the Literary Atlas of Europe. The data produced by such an approach would be available for use in contexts outside the academic realm including use in literary tourism.

4. Conclusions

This paper tentatively indicates that automated processing of text may support a procedural, iteratively based approach to geoparsing Ulysses that combines the application of software with manually identified terms. The project strongly suggests that the NER-CRF software was most effective in identifying explicit toponyms that were marked in the encoding as either routes or projected spaces. What the results of the project suggest also is that the application of typography might be partially automated; types of place may be determined through automated processes.

The application of the Literary Atlas of Europe's five categories of spatial representation as a place type within the encoding of the episode clearly supports the contentions of Gunn, Hart and others that place plays a dominant role in the Wandering Rocks episode.¹¹ The relative dominance of places of type "route" in the episode is not surprising and supports the notion of place as being significant to the novel.¹² What this approach indicates though is that such critical insights may be

verified using the algorithmic or distant reading frameworks. In this case, place is important to the episode because the majority of toponyms indicate explicit routes in particular, verifiable places.

However, one is also left with the insight that uncertainty, as marked by the absence of geographical identifiers, is the highest for certain types of places in the text. While place's significance to the novel is undoubtedly a likely outcome of a traditional, close-reading approach, the model proposed in the paper enforces a certain rigour in its approach to the text. Therefore, while the episode may be in some way explicitly "about" place, roughly a third of the places are of uncertain locations. This would markedly imply that, in the critical literature, Bulson's emphasis on the notion of disorientation and Hart's attention to the various "traps" of place can be traced back to measurable "quantities", within the confines of a constricted model, in the text. Additionally, one outcome of this approach is the difficulty in visually representing spatial uncertainty. This element is accommodated in the web component in terms of character routes rather than explicit location.

While toponym extraction and a geographically-contextualised approach towards the text enabled visual representations of the types of place in *Wandering Rocks*, the evocation of uncertainty, as facilitated by the LAE data model, made representation of such data challenging in a web-based, visual environment. The work described in this paper is generalizable within the larger Digital Humanities context as it demonstrates the practical application of NER, uses document encoding to explore meaningful geographic relationships in the text and leverages these relationships to interrogate spatial uncertainty.

References

1. **Budgen, F. & Hart, C.**, (1989). *James Joyce and the making of 'Ulysses' and other writings*, Oxford University Press.
2. **Bulson, E.**, (2011). 'Disorienting Dublin' in *Making Space in the Works of James Joyce* 1st ed. V. Benejam & J. Bishop, eds., Routledge.
3. **Moretti, F.**, (2007). *Graphs, Maps, Trees: Abstract Models for Literary History*, Verso.
4. **Clement, T.E.**, (2008). 'A thing not beginning and not ending': using digital tools to distant-read Gertrude Stein's *The Making of Americans*. *Literary and Linguistic Computing*, 23(3), pp.361 –381.
5. Anon, GeoNames. Available at: www.geonames.org [Accessed November 3, 2011].
6. **Edina**, (2009). *Unlock - Unlock Text: Geoparser Web service*. *Unlock - Unlock Text: Geoparser Web service*. Available at: unlock.edina.ac.uk [Accessed November 4, 2011].
7. **Bird, Stephen**, (2013). *Natural Language Toolkit*, Available at: www.nltk.org [Accessed March 6, 2013].
8. **Stanford Natural Language Processing Group**, (2013). *The Stanford NLP (Natural Language Processing) Group*, Available at: nlp.stanford.edu/software/CRF-NER.shtml [Accessed August 10, 2013].
9. **Piatti, B.**, (2013). *A Literary Atlas of Europe*. Available at: www.literaturatlas.eu/en [Accessed November 4, 2011].
10. **Hart, C.**, (1974). "Wandering Rocks" in *James Joyce's Ulysses*. Critical essays. Edited by Clive Hart and David Hayman., Berkeley: University of California Press.
11. **Gunn, I., Hart, C. & Beck, H.**, (2004). *James Joyce's Dublin: A Topographical Guide to the Dublin of Ulysses: with 121 Illustrations*, Thames & Hudson, Limited.
12. **Bulson, E.**, (2001). *Joyce's Geodesy*. *Journal of Modern Literature*, 25(2), pp.80–96.

DH on the Fringes: Using Smartphones, Instagram, and Ruby on Rails to Archive the DH Experience at an HBCU

Dighton, Desiree

desireedighton@gmail.com

Shaw University

Norberg, Brian

brnorber@ncsu.edu

North Carolina State University Libraries

Over the past few years, the literature on race in digital humanities has steadily grown. From various articles in Debates in the Digital Humanities to the development of Postcolonial Digital Humanities and MLA E-Roundtable, "Assessing Race in Digital Humanities", many have explored the theoretical and activist potential of addressing race in DH. Simultaneously, venues and projects like THATCamp and The Praxis Program have progressively pushed DH beyond the bounds of research one institutes. However, much has yet to be said about the complexities of involving students with digital humanities at under-resourced institutions. Teaching humanities classes at Shaw University, the first historically black college in the South, has given me an excellent opportunity to do just that.

This paper discusses how I dealt with a lack of technology and student confidence to create a modern archive of Shaw University student life, called #myshawu , by using smartphones and an open-source Ruby On Rails engine for harvesting Instagram photos. Though the assignment taught me the potential of DH as an empowering tool for my students to tell their stories in a public venue, it ultimately convinced me that the greatest hurdle to getting people of race involved in the field is not just exposure to technology, but teaching them the skills and critical thinking necessary for true digital empowerment.

The idea for the assignment began when I read the MacArthur Foundation report, "The Future of Learning Institutions in a Digital Age."¹ In the report, Cathy Davidson and David Goldberg disclose a sobering reality: "Despite government pronouncements to the contrary, 'digital divide' is not just an old concept but a current reality" (Davidson and Goldberg 20). The vast public acceptance of a so-called Internet Generation or Generation Y ignores the very real tech fluency differences that often exist along class and race lines. Siva Vaidyanathan points out just how detrimental this assumption is for underprivileged students in his essay, "Generational Myth": "to assume an entire generation is 'born digital' willfully ignores the vast range of skills, knowledge, and experience of many segments of society. It ignores the needs and perspectives of those young people who are not socially or financially privileged. It presumes a level playing field and equal access to time, knowledge, skills, and technologies."²

My experience teaching at one of the country's first HBCUs confirms the cautionary words of Davidson, Goldberg, and Vaidyanathan. Additionally, when I came across Adeline Koh's "Race and Digital Humanities: An Introduction" HASTAC presentation, Alan Liu's "Where is the Cultural Criticism in Digital Humanities", and other articles addressing the issue of DH and race, I began to wonder what digital humanities could do for my students. Could a DH project be a way to bridge this digital divide? Could a DH project be done with such limited resources?

Not only do many of my students lack technological fluency, many of them don't own computers, there are few computers labs on Shaw's campus, and the university has no technology unit that students or myself can turn to for help. My struggle with these limitations coincided with an interesting project being done at a neighboring institution. NCSU Libraries was using mobile devices to have students generate an archive of its new library in a project called my #hunlibrary. A few conversations and a partnership later, #myshawu was born.

#myshawu is a repository of student-generated images collected from Shaw's first-year, almost exclusively first-

generation students during a one-semester basic writing course into which 90% of Shaw's incoming freshman are placed. We bridged some digital access hurdles by using largely free tools and open source software. Students were already utilizing Instagram, an app accessible to most of them through their smartphones, a familiarity which lessened the learning curve and gave students the confidence to dive into the assignment. To gather all these images, I spun up an app which uses Lentil, the Rails engine that drives My #hunlibrary, on a free hosting service, Heroku. All the students had to do was use their Instagram accounts, take photos of their lives around campus, and tag them "#myshawu". Through these simple steps, students were able to represent themselves and share their images with other students, the university community, their communities of origin, and the public. As a companion to the photo collection, students published long-form, photo-rich narratives on a class collaborative Wordpress blog of the same title, # myshawu. These essays focused on their photos and their understanding of themselves in relation to the university and, hence, in relation to their academic identity.



Fig. 1: A page of student-generated images from the #myshawu website.

When students submitted their Wordpress essays, they also completed a survey to indicate the level of difficulty and the level enjoyment they experienced with the assignment. Out of 91 students who completed the assignment, 14 had difficulty getting the required hardware, 10 had trouble setting up required software accounts, and 15 students had problems using this software. These numbers show the relative ease with which students were able to access the necessary technology, an ease dependent on using personal and open-source tech resources rather than non-existent institutional tech resources. Yet, many of the students struggled to complete the assignment as successfully as I'd hoped. Only about half the images that students used in their final essays on Wordpress ended up on #myshawu, and a good many of these images were copied from the Internet. However, 85% of the students who completed the assignment said it was not difficult to take pictures.

There could be various reasons for this discrepancy between confidence and successful execution, such as poor resiliency or effort, and/or inconsistent classroom attendance. Yet despite some shortcomings, the unintended, often unquantifiable successes continue to emerge. As Shaw approaches its sesquicentennial and HBCU across the South consider closing their doors, the students involved with #myshawu are going to capture oral histories, chronicling in photos, video, and audio the stories of alumni so important to Shaw's history. The project has engaged many students beyond the classroom as they visit local history museums and inquire about the possibilities of exhibiting our work. They reimagine the project with me and ahead of me, asking for opportunities to use their pictures to raise funds to for my technology resources and for more campus events.

Another facet of this project shines a light on students' perspective on social media and adds another layer of complexity to the argument that media studies and DH scholars have been posing about race and technology. Logan Hill, in "Beyond Access", put it best: "Universal access isn't just about being able to surf the Web, it's about the ability to participate and compete in a technology-driven industry and society" (29).³ To have a wide range of students from all race and class backgrounds succeed in this media-saturated society, we can't just give them new technologies to create content only their friends will consume. We need to show them that the

technologies they use every day can be harnessed to empower their academic and professional lives.

#myshawu has also taught me the importance of digital humanities for helping my students develop a more critical perspective on technology. The full scope of what these kinds of projects teaches us about DH, HBCU students, cultural representation and empowerment is still unfolding, but it's clear through collecting and analyzing data, student-generated reflective writing, and evaluations, that digital tools like these cause necessary shifts in students' understanding of their own agency, particularly the role that writing, technology, and image creation can have on their power as scholars and professionals. Unfortunately, largely due to financial constraints and the many needs pulling at institutions like Shaw, HBCUs are some of the most unlikely to support faculty technology training and least likely to have the resources to support digital projects. Yet DH tools belong in the hands of those who have the most at stake as they become invaluable tools for engagement and student success. That said, certain assistance is needed from the digital humanities community for these projects and these students to reach their full potential. Partnerships with other universities, academic technologists, and more accessible and flexible technologies are just some of the support that can be extended by the greater DH community to HBCUs like Shaw, thereby creating cross-institutional collaborations that highlight the strengths of all institutions. Then #myshawu can become just one effort in a larger movement towards developing a DH pedagogy that can be inclusively practiced without relying on the financial privilege so often synonymous with DH.

References

1. Davidson, Cathy N. and David Theo Goldberg (2009). *The future of learning institutions in a digital age*. Cambridge, Mass: MIT Press.
2. Vaidhyanathan, S. (2008). *Generational myth: Not all young people are tech-savvy*. The Chronicle Review, 55(4), B7. Retrieved October 19, 2013 from: chronicle.com.
3. Nelson, Alondra and Thuy Linh N. Tu with Alicia Hedlam Hines (2001). *Technicolor: Race Technology, and Everyday Life*. New York: New York University Press.

Exploring the Intersection of Personal and Public Authorial Voice in the Works of Willa Cather

Dimmit, Laura

lkdimmit@gmail.com

Nebraska Literary Lab, University of Nebraska-Lincoln

Kirilloff, Gabrielle

gkirilloff@gmail.com

Nebraska Literary Lab, University of Nebraska-Lincoln

Warren, Chandler

chandlerawarren@gmail.com

Nebraska Literary Lab, University of Nebraska-Lincoln

Wehrwein, James

jawehrwein@gmail.com

Nebraska Literary Lab, University of Nebraska-Lincoln

1. Introduction

In a letter that Willa Cather wrote in response to a reader's critique in 1924, she claims: "I had a perfectly good reason for writing 'Antonia' in the first person, masculine—and I did not for one minute try to 'talk like a man'. Such a thing as humbugging any one never occurred to me. It does not matter who tells a story. It is merely a point of view, a position which the writer takes in regard to his material..."¹

Looking into an author's unpublished writings for additional insight is an analytic strategy with a long history in literature studies. With selections from Willa Cather's extensive collection of correspondences now available for study, scholars are presented with the opportunity to compare Cather's private letters with her published novels. A statistical examination of Cather's letters has, until recently, been impossible; until the 2013 publication of *The Selected Letters of Willa Cather*, Cather's letters were not available to the public.² The publication of the letters offers a glimpse into the private life of Cather, a life that she kept carefully guarded. The collection contains approximately 550 letters, spanning her life from age 14 until just days before her death. While Cather often spent years working on her published writings, the fact that many of her correspondences were hurriedly dashed off indicates that an examination of the letters could reveal a style of writing that is less polished, worked, and intentional.

The potential applications of this type of statistical analytic in literature studies as a whole are wide reaching. Many authors have left behind extensive collections of personal writing that could hold untapped insight about how they communicated in different settings, and how their writing evolved. By using similar tools to those we utilized in this exploratory research of Cather, scholars can continue to quantify the suppositions that traditional methods of literary analysis have yielded.

2. Methodology

Our research was divided into two main tasks. First, we established a "personal voice signal" from Cather's correspondence. To derive a quantitative signal, we narrowed the available letters down to 164 letters using the hclust() function in R (For additional information, see <http://www.r-project.org/>). Setting the number of clusters to 18, we chose the cluster that contained the highest percentage of letters addressed to distinctly personal correspondents. The personal signal was created by using a mean frequency threshold to determine a set of frequently used words in the corpus of letters. In the tradition of Burrows and others working in authorship attribution, we elected to measure style based on the occurrence of high frequency word features^{3 4}. This allowed us to both avoid arbitrarily selecting words, and to avoid using any context-sensitive words in our signal.⁵ It is likely that the 15 frequency features, or function words, that we selected are also less dependent on an author's conscious decision to vary his or her vocabulary.⁶ Thus, function words might offer a better indication of the subconscious, intrinsic choices that inform a writer's voice. It was our initial expectation that Cather's personal voice signal would differ from the narrative voices present within her novels. Our initial research question revolved around the relationship between private and public authorship: how (if at all) did Cather's fiction act as a mouthpiece for her personal voice?

Once we generated a signal from Cather's letters, we used a clustering function in R to compare the signal with 15 of her novels. In order to compare Cather's novels with the personal voice signal derived from her correspondences, we used R to calculate two statistical measures of similarity: correlation coefficient and Euclidean distance. These two units of measure compare the similarity between Cather's personal voice signal and her use of function words in the corpus of novels. Examining the Euclidean distance (with the dist() function) provides a way to calculate the numerical distance between the frequency of the function words within the personal correspondences and the frequency of the same set of function words within each of the five novels. The correlation coefficient measures the linear dependence between the function words in the correspondences and the function words in each novel. We divided each of the novels into 1,000 word chunks in order to track the extent to which portions of the novels were more similar to Cather's personal voice than others. In order to calculate a unique Euclidean distance for each novel, as opposed to its parts, we averaged the distances for all the 1,000 word sections of each novel to determine a mean similarity between each novel and Cather's personal voice signal.

3. Observations

While some of our findings on the similarity between Cather's personal voice signal and her use of function words within her novels reaffirmed our initial hypotheses, some have defied initial expectations. We found that Cather's use of function words within her novels was significantly different from the personal voice signal calculated from the correspondences. However, the frequency of function words did not vary significantly among 14 of the 15 novels we examined, regardless of the gender of the narrator. This would indicate that, while there is a difference between Cather's "personal voice" and 14 of her published novels, certain aspects of the voice she adopts in these novels remains constant regardless of the identity of the narrator. We did identify a single outlier among Cather's novels. The frequency of function words within Cather's novel *My Mortal Enemy* more closely resembles the use of frequency words in Cather's correspondences than any of the other novels we examined.⁷

My Mortal Enemy was published in 1926, approximately halfway through Cather's publishing career. Previous scholarly commentary on Cather has also indicated that *My Mortal Enemy* is a singular book in her fiction corpus. The novel focuses on two periods in the life of Myra Henshaw and her husband, Oswald. The narrator, Nellie Birdseye, who many scholars argue is modeled after Cather, acts primarily as a frame for sharing the Henshaw's story. It is widely considered to be the embodiment of Cather's own novelistic ideal, the novel demeuble.⁸ For Cather, the novel demeuble was an evocative form of realism, stripped of unnecessary detail and embellishment. The unique style of narration present in *My Mortal Enemy* raises interesting questions about similarities between this novel and Cather's correspondences. If *My Mortal Enemy* embodies Cather's own idealized notion of stripped bare, to the point realism, than is it possible that Cather also adopted a similar style in her personal letters?

Interesting questions are also raised by the way in which *My Mortal Enemy* draws from Cather's life. The novel's autobiographical nature has largely been shrouded in mystery since Cather carefully attempted to conceal the events and people that influenced her novels. Charles Johanningsmeier has made one of the most thorough attempts at unraveling the connections, tracing the novel's influences to Cather's history with the McClure family.⁹ In this sense, our work appears to support Cather studies by backing up scholarly claims about autobiographical influence with quantitative data. We believe that examining the relative similarity or difference between the 'voice' of Cather's letters and fiction is relevant because of the claims she made. Based on comments made in some letters, it appears that she perceived differences in her writing that surpass the conventions of genre. If we assume that in writing about deeply personal issues, an author is likely to lapse into a more personal style, our findings seem to support the claim that *My Mortal Enemy* was highly influenced by Cather's personal life. This raises an interesting question: does the use of autobiographical details in Cather's fiction indicate an unconscious use of a more personal style of writing?

4. Future Work

In the future, we intend to deepen and expand the traits that define the personal voice signal we are deriving from Cather's correspondences. Sentence length, structure, and the choice of infrequently appearing vocabulary are all elements of a writer's unique voice. Our research team will be investigating the ways in which these characteristics impact the relationship between Cather's private letters and her published works.

It is also our hope that the research we have begun will lead towards further studies on the relationship between *My Mortal Enemy* and Cather's letters. In addition, since scholars recognize the main character from *My Mortal Enemy*, Nellie, as modeled after Cather, future research might begin to look specifically at the ways in which Cather's personal voice is associated with specific characters in her novels. This question might again return to issues of gender; if it becomes clear that

certain characters within Cather's work speak in a way similar to Cather's personal voice, the next question should revolve around what these characters have in common.

References

1. **Cather, Willa.** (2013) *To Mr. Miller. The Selected Letters of Willa Cather.* Ed. Andrew Jewell and Janis Stout. New York: Alfred A. Knopf. 362.
2. **A. Knopf** (2013). *The Selected Letters of Willa Cather.* Jewell Andrew, and Janis Stout, eds. *The Selected Letters of Willa Cather.* New York: Alfred.
3. **Grieve, J.** (2007). *Quantitative authorship attribution: an evaluation of techniques.* Literary and Linguistic Computing: Journal of the Association for Literary and Linguistic Computing, 22(3): 251–70.
4. **Burrows, J.** (2002). *'Delta': a measure of stylistic difference and a guide to likely authorship.* Literary and Linguistic Computing: Journal of the Association for Literary and Linguistic Computing, 17(3): 267–87.
5. **Jockers, Matthew L., Daniela M. Witten, and Craig S. Criddle.** *Reassessing Authorship of the Book of Mormon Using Delta and Nearest Shrunken Centroid Classification.* Literary and Linguistic Computing, 23.4 (2008): 465–492.
6. **Ray Siemens, John Unsworth** (2004). *A Companion to Digital Humanities*, ed. Susan Schreibman. Oxford: Blackwell.
7. **Cather, Willa** (1926). *My Mortal Enemy.* New York: Alfred A. Knopf.
8. **Cather, Willa.** (1922) *The Novel Demeuble.* The New Republic, 30.1: 5–6.
9. **Johanningsmeier, Charles.** (2003) *Unmasking Willa Cather's 'My Mortal Enemy.'* Cather Studies, Volume 5: Willa Cather's Ecological Imagination. n. pag. Web.

Representation and Absence in Digital Resources: The Case of Europeana Newspapers

Dunning , Alastair

The European Library, Netherlands

Neudecker, Clemens

The KB National Library of the Netherlands, Netherlands

Within the Digital Humanities, there is a long history of debate and discussion as to how texts are accurately represented in digital form. Arguments as to how texts are encoded in both a logical and semantic sense are a recurring feature of past DH conferences.

Yet the intense intellectual focus on the precise details of marking up small corpora or even individual texts has masked the fact that issues related to the representation of large corpora of digitised materials - books, manuscripts, newspapers, records etc. - have been too often ignored. Libraries, archives, museums and other collection institutions have now been digitising corpora of material for many years, but with a very few exceptions, it is still quite rare for an entire run of primary sources to be digitised and made available online.

This means that there are gaps within the digital record. Yet it is unusual for online resources to actively demonstrate these gaps; resources may be advertised as a growing corpus, but when searching through or downloading a digital resource there is rarely any indication of what has not been digitised. This skews the sense of the nature of the collection the scholar is working with and erodes trust.

This problem is compounded by assumptions made by end users that when a search is made in a digital resource, they actually are searching over everything in the original archive. In most cases, this is far from being the case.

This long paper looks at this problem in the context of the Europeana Newspapers project (www.europeana-newspapers.eu)

, a three year, four million euro project, which is creating full-text for 10m pages of digitised newspapers from 12 libraries across Europe, and also developing an interface to allow for cross searching of over 18m newspaper pages. The final interface, available from the European Library in 2014 (www.theeuropeanlibrary.org), will also provide keyword searching over the OCRd (Optical Character Recognition) text and allow users to compare different newspapers from around Europe published on the same day.

While it is an ambitious project, it is only a drop in the ocean of the overall number of digitised newspapers in Europe (a conservative calculation within the project put the number of digitised newspaper pages in European libraries at 130m). What appears on the final interface will only be a sample of what actually exists in European libraries.

Moreover, other issues - political, economic, legal and technical - mean that the quality and national distribution of newspapers in the project (and therefore represented in the final online interface) are unevenly balanced. For the resource to be trusted by the academic community, this lack of balance must be acknowledged

In terms of the economic and legal issues, the project is integrating newspapers from 12 existing newspapers online libraries, each of which have different business models. These different business models affect the final project interface. The National Library of Turkey and the British Library newspapers operate behind a pay wall, for instance - therefore the final Europeana Newspapers site will not be able to directly show images from their collection.

Other libraries are wary of sharing full-resolution images, with the legitimate fear that the users will no longer visit their own national website. In such cases, only fragments of their newspaper images will appear in the central site. Legal issues are also pertinent; some libraries are unsure of the copyright status of some of their historic newspapers and therefore do not want to commit to allowing another entity to publish them

In addition, there are several technical issues impeding uniform access to the resources. Nearly every digital newspaper collection today contains full-text derived from automatic processing with OCR software. But while some newspaper repositories grant access to the full-text, often the full-text is hidden and only exposed as an index for searching, but not available to the end user for online display or (programmatic) download, or sometimes not even for indexing by Google.

In other cases, full-text is made available, but not for the entirety of the collection, either due to IP issues or because the content holder took a deliberate decision not to show the full-text to the user, often because of the amount of error rate in the OCRd text. Regularly there is no sufficient information provided about the OCR error rate of a particular digital resource, which makes it even harder to assess what amount of the content can realistically be retrieved through a full-text search.

There are also different ways how digital facsimiles are made accessible. Many recent online newspaper portals use the JPEG2000 image file format. The benefit of this is the ability to zoom more or less seamlessly in and out of the digital facsimile. But since JPEG2000 has not been around for a very long time in the digitisation community, many collections that have been digitised in the past are only available in TIF format. This means that zooming can only be provided in a static way on these images, e.g. through different resolution JPEGs. As a result, it is often not possible for researchers to explore these legacy resources in much the same way as they do with recently digitised materials.

In other cases, digital facsimiles have been produced by capturing existing microfilm copies rather than the original source material, thus the digital versions expose artefacts that were not present in the original paper source, but only introduced in the microfilm. However, this type of provenance is most typically not available to end users who are left alone in their interpretation of the differences in resource presentation and functionality.

Finally, the metadata standards used to describe the digital contents also vary. Not only are there different representations in use for encoding full-text such as plain text, ALTO or TEI. But also descriptive metadata is commonly encoded in different

standards, and with different degrees of granularity. While standard bibliographic information such as the title or date of publication are commonly available, more specific information on, for example, a particular article or the names of persons or places occurring in it rarely are. Within the Europeana Newspapers project a subset of 2m pages out of the total 10m will be refined further down to the article level, thus enabling more sophisticated search and retrieval functionality than the remaining 8m pages.

A central point of this paper is that these issues are not just issues for librarians; it is not about showcasing how a digital resource is. Rather it is the urgent need to demonstrate how such issues have a profound effect on the academic community's engagement with online resources.

If a researcher wants to conduct a comparative analysis of newspapers in Chronicling America (the US historic newspaper site), the National Library of France and the British Library, she will have to use three different interfaces with different levels of content and metadata quality. Moreover, she will also have to grasp the particularities of each of these collections with regard to their quality and completeness and what that entails for her research.

This paper will conclude with some recommendations for how those building digital resources can make their content choices more transparent. Informed dialogue between the cultural heritage organisations and the research communities is required. It calls for creators to tear down the illusion of completeness and help persuade end users that many digital resources are fragmentary things, where the representation of absence is just as important as representation of existence.

References

For a brief summary of the issue see **Julia Flanders**, "Collaboration and dissent: challenges of collaborative standards for digital humanities" in Collaborative Research in the Digital Humanities (eds. Marilyn Deegan, Willard McCarty), 2012. The TEI mailing list provides ample evidence of such discussion listserv.brown.edu/archives/cgi-bin/wa?A1=ind1309&L=TEI-L.

For instance, **Johanna Drucker** in "Performative Materiality and Theoretical Approaches to Interface" Digital Humanities Quarterly (2013, Volume 7 Number 1) and also "Humanities Approaches to Graphical Display" Digital Humanities Quarterly (2011, Volume 7 Number 1) addresses theoretical concerns relating to the interface but with less focus on its practical representation within online resources. The issue has received much more attention in the world of 3D visualisation, e.g. with the creation of the London Charter (www.londoncharter.org).

See "History, Digitized (and abridged)" for a summary of the extent of digitisation in 2007. www.nytimes.com/2007/03/10/business/yourmoney/11archive.html?pagewanted=all&_r=1.

One of the findings in, *Reinventing research? Information practices in the humanities*, Research Information Network, 2011, www.rin.ac.uk/our-work/using-and-accessing-information-resources/information-use-case-studies-humanities.

David Nicholas, Ian Rowlands (2007), *Google Generation*, Paul Huntington www.jisc.ac.uk/whatwedo/programmes/resource/discovery/googlegen.aspx.

Alastair Dunning, European Newspaper Survey Report, 2012, www.europeana-newspapers.eu/wp-content/uploads/2012/04/D4.1-Europeana-newspapers-survey-report.pdf.

For a comparative study of search ranking of digital newspaper repositories see *Digital collections: If you build them, will they visit?*, **Frederick Zarndt et. al.**, IFLA WLC2013, Newspaper and Genealogy Section, Singapore, www.ifla.org/files/assets/newspapers/Singapore_2013_papers/day_1_01_xzarndt_frederick_et_al_digital_collections.pdf.

Jan Hillgärtner (2013), *Digitalisierte Zeitungen und OCR: Welche Forschungszugänge erlauben die digitalen Bestände?*, 18/03/13, newsphist.hypotheses.org/23.

For a study in the methodology and analysis of digitised newspapers vs. paper copies see *The Digital Turn. Exploring the methodological possibilities of digital newspaper archives*, **Bob Nicholson**, in Media History Vol. 13, Issue 1 2013, Special issue: Journalism and History: Dialogues.

For an example of this issue within the Digging into Data projects see *One Culture. Computationally Intensive Research in the Humanities and Social Sciences A Report on the Experiences of First Respondents to the Digging Into Data Challenge*, **Christa Williford and Charles Henry**, (2012) www.clir.org/pubs/reports/pub151 and also the aforementioned Reinventing research? Information practices in the humanities.

On Reusability and Electronic Literature

Durity, Anthony

University College Cork, Ireland

O'Sullivan, James

josullivan.c@gmail.com

University College Cork, Ireland

1 Writing Longevity

This paper is a return to an open argument: it addresses a critical issue that, as of yet, remains unanswered. Through direct engagement with notable digital writers, we explore issues of reusability and obsolescence in electronic literature. A historical account suggests that closed formats seemingly dominate the electronic literary landscape, contrary to the open culture which writers, publishers and scholars working within this field tend to promote. In order to investigate this matter, we focus our attention on two of the field's most prominent anthologies, the Electronic Literature Collection, Volume 1 (2006)[hay, 2006] and Volume 2 (2011)[bor, 2011], curated by the Electronic Literature Organization. Of the 62 works anthologised in Volume 1, an assessment of the 25 pieces considered poems shows that only five can be thought of as open. By "open", we mean that readers can reasonably access the underlying code. By "open" we do not mean to refer to open access publishing. Our paper is about reusability and access at the level of content creation not consumption, open-source not open access. Thus, in the earlier collection, approximately 20% use an open format. In Volume 2, over half the works use closed formats. By availing of closed-source approaches to publication, the reusability of digital literature is sacrificed, and the works remain at the risk of obsolescence. Furthermore, the open ethos of digital culture is neglected. Non-digital poetry is presented in an open format. The codex is open, developed heuristically over centuries courtesy of a speculative problem-solving feedback loop between author and publisher. The transcribed word has escaped from the monk's cell. The only time that language is obfuscated is when the addresser wishes to keep the contents of a communication understandable or decodable by a select few. Contrary to the common view that digital art is more open than its predecessors, electronic literature is actually at variance with centuries of open book culture. We are not concerned with commercial or proprietary considerations, but with the notion of "openness". We do not deny the disseminative qualities of electronic platforms, but focus solely on the openness of the cultural apparatus. Our purpose is not to criticise the ELO or its authors, but to determine why they chose closed platforms, and if issues surrounding reusability, digital preservation, production and maintenance costs were factored into their creative decisions. To their credit, the ELO and the field's leading scholars recognised these issues from creative literary practices: "Electronic literature doesn't come on bound, offset-printed pages. Keeping it on a shelf doesn't mean that it will be easy, or even possible, to read it in the future. Even putting it into a vault with controlled temperature, light, and humidity won't ensure its availability." [Montfort and Wardrip-Fruin, 2004] In recommending approaches to the preservation of electronic literature, Montfort and Wardrip-Fruin encourage authors to

avail of open standards: "Those who use open systems and adhere to open standards when creating electronic literature have a much better chance that the format of their literary works will be supported, or decipherable, in the future." [Montfort and Wardrip-Fruin, 2004] Their warning to authors is based on the reality that closed systems and unknown specifications "are far more difficult to migrate and emulate", and that such systems are typically controlled by small groups which "may lose interest" or change a "standard without warning, so that older works of electronic literature no longer work on new platforms." [Montfort and Wardrip-Fruin, 2004] However, Montfort and Wardrip-Fruin also acknowledge that authors do base their selections on artistic considerations: "A closed system may provide important capabilities that are otherwise not available, and some closed systems may be very well suited for the type of literary creation in which authors are interested, so there may be good reasons for authors to use a particular closed system." [Montfort and Wardrip-Fruin, 2004] Authors doing so must be conscious, they argue, that "such a choice could affect the longevity of their works" [Montfort and Wardrip-Fruin, 2004]. Montfort and Wardrip-Fruin's "Acid-Free Bits: Recommendations for Long Lasting Electronic Literature" is the seminal account of how digital literature should be developed with the threat of obsolescence in mind. In this paper, we hope to build on their work, identifying if authors are indeed mindful of such recommendations, and what precisely influences their decisions when it comes to choosing a platform.

2 Methodology

We surveyed contributors from both volumes of the Electronic Literature Collection since this provided a list of authors whose work is considered, by the field's most respected body, of a standard suited to publication under the mantle of "digital literature". This allowed us to avail of convenience sampling. We developed a brief questionnaire comprised of open-ended questions, allowing respondents the freedom to provide answers that were not shaped or guided by our assumptions. The questions were as follows: 1. When creating those poem(s) included in the ELO Collection, what were your reasons for choosing the technologies that you did? 2. As a writer/artist working with digital media, do you take into account issues relating to reusability? 3. As a writer/artist working with digital media, do you take into account issues relating to obsolescence? 4. As a writer/artist working with digital media, do you take into account issues relating to production and maintenance costs? 5. On the subject of technologies (software and/or hardware) being adopted by digital writers/artists, is there anything else that you would like to add? 6. If you are interested in engaging with us further in relation to your work as a digital author, please provide your name and preferred contact details. You can be assured of anonymity unless otherwise agreed. The research question did not require a measurement or comparison of groups, rather, it sought to elicit the technological motivations of authors. A pluralistic approach to qualitative analysis was used as we were reluctant to be constricted by any one method, potentially missing the importance of certain passages. Repeated readings of the data, combined with a comparison of coding, meant that a more complete understanding could be achieved.

3 Findings

A number of thematic axes emerge from our interpretation. One corresponds to author notions about the perceived fragility or stability of digital platforms. Transience and fluidity have clearly come to be associated with digital modes of encapsulation. This is arguably reflected in the subject matter which many of these authors treat; the fragility of memory. As evidenced in our data, associations are made between the ease with which an artefact can be replicated, and the ease with which it might achieve permanence. Authors view the fluidity of digital media as opposed to the fixity of print. This runs counter to the true nature of some digital media, particularly closed standards. It is interesting that some authors make what is,

particularly in relation to the standards they adopt, something of a false association. It may be a product of authors falling victim to the hyperbolic language that surrounds digital platforms, or, this theme may have emerged because authors focus on the text, rather than the underlying electronic systems. This was supported by explicit responses: "I have always tried to [take into account issues relating to obsolescence]. And yet most of my works are dead now. I think that is the nature of the beast...Now I do only work in HTML/ASCII text so it can be resurrected more easily." To further test this hypothesis, this paper will present findings as to how digital authors view their work, and whether or not they make a distinction between the literary and paratextual elements of their pieces. Doing so is significant in discerning if author expectations match reality. Another theme to emerge is that of collaboration, which has become an important part of the literary process, particularly in relation to electronic works. This can dictate the material aspects of a piece, as authors are restricted to using familiar technologies. Furthermore, we suggest that a lack of reusability – in essence, simple building blocks – may be the cause of this reliance on collaboration. Respondents tended to cite personal reusability in favour of more communal approaches: several authors revealed that they have their own library of reusable components. Restrictions and constraints that are a product of technical expertise are somewhat unique to this medium, as writers are required to work with, rather than for, the medium. Authors of non-digital literature send their manuscripts to publishers, who in turn produce a book – the author does not necessarily know the workings of the production process. Literary production on digital platforms requires that authors possess at least an understanding of computing sufficient for communication with their technical collaborators. We are going to present the findings of our interactions with the authors rather than our perspectives – therein lies the novelty.

4 Pragmatism & Virtue

Software development itself can be viewed ethically or pragmatically. That is, some would declare that software development done in the open is measurably better in a practical sense. Others would hold that it is ideally better in the moral sense. Electronic literature may use open or closed formats for distribution. The classic modern examples of this are Web technologies versus products like Adobe Flash. The fact remains that authors may still choose to distribute the source-code for their creations even if the packaging is closed. Furthermore, just because the packaging of the Web is open does not mean that the author has made any concessions to reuse and reusability. Few authors here have placed their materials and code into online repositories and/or code versioning systems. We take both approaches to our assessment. Pragmatically, successful digital phenomena like the Internet, built on open protocols, and Android, built on open source, indicate that this model serves authors and innovators best. Ethically, we take virtue theory and apply it to epistemology: considering the act of sharing from a value-neutral position, the virtue (the mean) is openness, the lack of which is secrecy, with "promiscuity" being excess. To contextualise this position, consider non-digital literature, and the level to which its constituent parts are invisible to its readers. We will present our findings to the Digital Humanities community, entering into discussion on the construction of digital literature as typified by the ELO's contributors, addressing the reusability of such from pragmatic, ethical and literary perspectives.

References

- (2006). *Electronic literature collection 1*. In Hayles, Montfort, Retberg, and Strickland, editors, *Electronic Literature Collection Volume One*.
- (2011). *Electronic literature collection 2*. In Borràs, Memmott, Raley, and Stefans, editors, *Electronic Literature Collection Volume Two*.

- Aarseth, E. J.** (1997). *Cybertext: Perspectives on Ergodic Literature*. The Johns Hopkins University Press. Book, Whole.
- Benjamin, W.** (1934). *The Work of Art in the Age of Mechanical Reproduction*. Penguin, London.
- Hayles, N. K.** (2007). *Electronic literature: What is it?* v1.0.
- Heidegger, M.** (2010). *The question concerning technology*. In Krell, D. F., editor, *Basic writings: Martin Heidegger*. Routledge, London.
- Hyde, L.** (2012). *Common as Air: Revolution, Art, and Ownership*. Union Books, London.
- Manovich, L.** (2002). *The Language of New Media*. The MIT Press. Book, Whole.
- Montfort, N. and Wardrip-Fruin, N.** (2004). *Acid-free bits: Recommendations for long-lasting electronic literature*. v1.0.
- Wardrip-fruin, N. and Montfort, N.**, editors (2003). *The New Media Reader*. The MIT Press.
- Williams, S.** (2009). *Free as in Freedom, volume paperback edition reprint*. SoHo Books, Lexington, KY, USA. Book, Whole.

Digital Yoknapatawpha: Interpreting a Palimpsest of Place

Dye, Dotty J.

Arizona State University, Tempe, Arizona

Napolin, Julie Beth

The New School for Liberal Arts, New York, New York

Cornell, Elizabeth

Fordham University, Bronx, New York

Martin, Worthy

University of Virginia, Charlottesville, Virginia

Digital Yoknapatawpha is a critical database, interactive map, timeline, and network visualization that aims to unite the 15 novels and 48 short stories that William Faulkner, over the course of his career, set in the imagined county of Yoknapatawpha. “Digital Yoknapatawpha: Interpreting a Palimpsest of Place” reports on the collaborative effort of 25 scholars from around the world, led by Stephen Railton, to expand the understanding of Faulkner’s creation and recreation of this one county. This paper, written jointly by four of the project’s collaborators, will describe the digital methods and practices engaged in the project, as well as the inherent challenges produced when digital spaces encounter the fictional world of an author who did not care much about getting the facts right.

William Faulkner, the 20th-century writer known for his portrayal of the American South, envisioned his fictional world, Yoknapatawpha County, as a series of maps. In Faulkner’s development of Yoknapatawpha’s geography through narrative, the creators of *Digital Yoknapatawpha* have discovered the patterns of a digital space, meaning that Faulkner returns to the same places, times, and people in his fiction, though he often alters the details. Like a digital space, these details are ever revisable. Yoknapatawpha County’s imagined geography branches out in a web of connections between people, places and events through multiple representations. Each element layers onto another, creating a palimpsest of place throughout Faulkner’s fictional world. For our purposes “place” is not just a point (lat/long) or even a simple area on a map. Certainly, geography is crucial, however, the “place” that constitutes our primary interest is established by the confluence of human activity at locations, across areas and reverberating in time. To be sure, we can gather ideas about Faulkner’s corpus without the use of digital methods. We argue, however, that without the meticulous harvesting and collation of these elements we can, at best, only derive vague impressions of the complicated webs that inhabit Faulkner’s fiction. While a substantial number of Faulknerian scholars already focus on the role of “place” in understanding human activity, our intent is to complement and extend such scholarship, not supplant it, through internet-accessible interactive visualizations.

Our collaboration began with discussions (recurring small-group meetings and a National Endowments for the Humanities funded, multi-day, full-group workshop) in which we worked to establish “editorial policies” that could guide the interpretive decisions that each team of editors would face as it conducted the close reading of individual stories and novels. Fundamental to the interpretation is the agreement that Faulkner’s “confluence of human activity” may be digitally interpreted by identifying the characters that participate in the events Faulkner composes and by identifying the locations and time-frames attached to each event.

Each team therefore conducts close readings of an individual text with the goal to draw out specific details about the categories of characters, events, locations and time frames as keyed to that specific text. The interpretations derived from those close readings form the basis of the interactive visualization we call *Digital Yoknapatawpha*. While the interactive aspect of the visualization cannot be conveyed here, we have included Figures 1-5 to provide a basic overview of the information available through the visualization.

The novel, Flags in the Dust, and 11 short stories that constitute the current set of interpreted texts are shown on the prototype of the home page in Figure 1. Selecting the leftmost icon takes one to the visualization specific to Flags in the Dust, as shown in Figure 2. The central area of the display depicts a map of the imagined geography engendered by that text. It contains man-made features (e.g., roads), natural features (e.g., hills and waterways) and the location icons that appear as settings for events in the novel. These are in layers that can be hidden or shown via the controls on the left. To the right are generalized geographies for events that might be interpreted as taking place outside of Yoknapatawpha. When a user selects a specific location on the map, the map yields a textual display that describes the location and provides tabs listing the events occurring at that location and the characters participating in those events (see Figure 3).

The dynamic mode of the visualization becomes crucial when we consider the progression of events in the narratives. We use the term “progression” to emphasize that events play out in two spacio-temporal modes: the narrative or textual order (indicated by page number) and the chronological order as we interpret it using historical and calendar clues within the text. The three bars on the bottom of the display (see Figure 2) provide playable timelines; the uppermost play buttons provide a textually organized progression and the lower buttons provide a chronological one.

Figure 4 provides a snapshot of the progression in which the second event has been selected on the page-order progression bar. Allowing the page-order to “play” through to an event that occurs on page 75 yields the display shown in Figure 5. In this mapping one can see the geographic expanse of the events that have occurred up to that point in the novel while the chronological line demonstrates the way that the novel has narrated human activity across various temporal settings.

As presented above, in the *Digital Yoknapatawpha* project, we use digital methods that both challenge and mimic Faulkner’s aesthetic project in order to discover new insights about his narrative approach. For example, Faulkner manipulates, to great effect, narrative space against textual space. By entering events with a page-order aspect, we are forced to focus on the textual space as Faulkner deliberately shaped it. When we order those events into a chronology—a chronology that perhaps spans his entire fiction—we are forced to confront, in very concrete ways, the absences, contradictions, concurrences, and shifting perspectives that spread out over that textual space.

So much in Faulkner is speculated, projected, imagined, and contradicted. Faulkner himself often claimed to be interested in truth, not facts. Therefore, how can we account for the ambiguity evinced through close reading yet also record the specific information required for visualization? The specific manner of close reading that our project of data collection requires forces us to identify, yet at times bracket, the narratives’ ambiguities and contradictions in order to determine precise locations in the textual space. This approach, and its attendant frustrations, allow us to enliven Faulknerian contradiction with traditional maps and innovative network

visualizations, thus making the ambiguity that is so central to Faulkner's fiction also a central element of the project's design and display. In turn, the potential in the aesthetics of digital space overcome the challenges and frustration of contradictory positions by providing visual and interactive ways to engage with ambiguity in Faulkner's work.



Fig. 1: Home page indicating novel and stories covered.

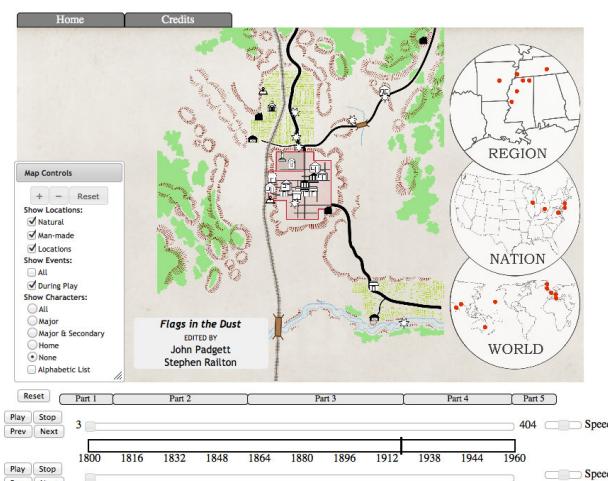


Fig. 2: A visualization of the imagined geography (and associated geographies).

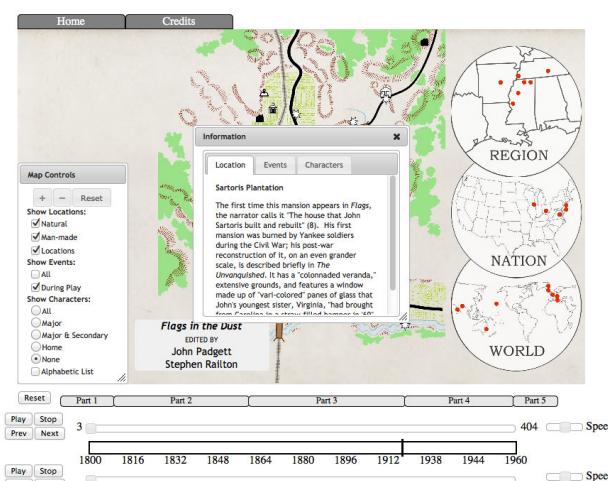


Fig. 3: Textual information about a specific location.

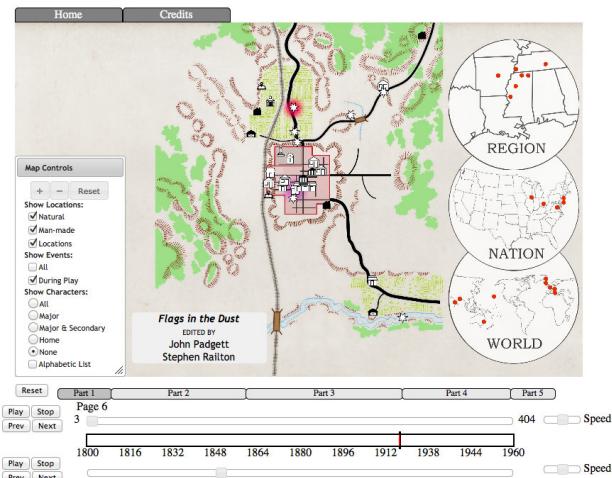


Fig. 4: A location highlighted to indicate an event is happening there.

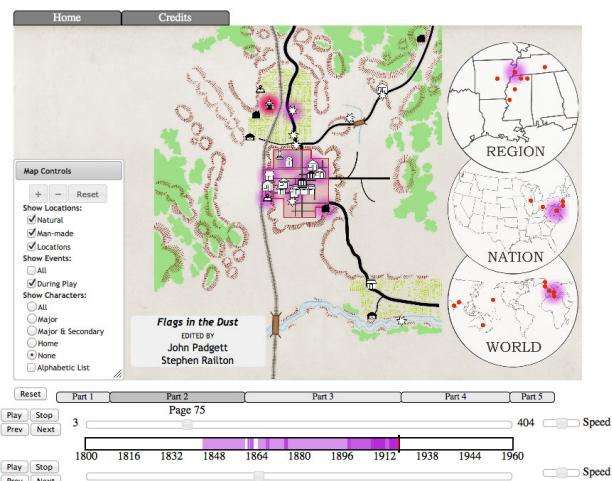


Fig. 5: A different location highlighted, with the accumulated evidence of events already told.

References

See: <http://faulkner.iath.virginia.edu/prototype/>

Digital Activism: Canon Expansion and Textual Recovery in the Undergraduate Classroom

Earhart, Amy

aearhart@tamu.edu

Texas A&M University

Taylor, Toniesha

ttaylor@PVAMU.EDU

Prairie View A&M University

In Earhart's recent essay "Can Information Be Unfettered?

Race and the New Digital Humanities Canon,"¹ Earhart critiques the digital "canon that skews toward traditional texts and excludes crucial work by women, people of colors, and the GLBTQ community" (316), advocating an activist model of grassroots recovery projects to expand current digital offerings.

In response to such concerns, Earhart and Taylor are currently testing a digital recovery project entitled White Violence and Black Resistance in Texas. This year's conference theme is Digital Cultural Empowerment, a theme directly related to our testbed project which emphasizes the expansion of cultural capital and digital literary skills through our model pedagogical project. We view our project as connected to interventions

into current structures of production through the digitization and dissemination of materials about white violence and black resistance found buried in difficult to access rare book rooms, crumbling newspapers, analog and/or transcribed oral histories, and unknown journals. Working with primary materials of the reconstruction period, the project seeks to illuminate a time of great cultural conflict within Texas. Newspapers, Freedmen's Bureau records, marriage records, census records, legal records and oral histories illuminate the response to and resistance of emancipated African-Americans from 1867 through the turn of the century. Other projects, such as Black Gotham, have made such archives open to the public, but no project to date has been modeled on the type of student research learning model we are utilizing. Our ongoing project presents an activist model grounded in the classroom where undergraduate students are participants in canon expansion while learning valuable research and digital literary skills, a model that we believe other digital humanists interested in canon expansion and digital pedagogies might replicate.

Our project is developed to answer two challenges in current digital humanities practice. First, we view our project as a way to leverage expertise and resources across historical areas of divide. Criticism of current institutional digital humanities practice has targeted the divide between institutional haves and have nots, often portrayed in tensions between well funded research institutions versus small teaching institutions. Our testbed of Texas A&M University and Prairie View A&M University² provides an important site of intervention. Founded in 1876, the two universities were divided by race, during segregation, and finance. Though the state constitution in Texas states that both are "universities of the first class" they have not seen funding and resources that make this true. Rather, the campuses have continued to be marked by the separations of race and resources which creates a space where Texas A&M is constructed as a predominantly white research university and Prairie View a historically black teaching university. We reject the differentiations and view projects like our current work as a way of disrupting such binaries, using carefully constructed technological projects to spread digital cultural empowerment through both universities and student bodies. Here we agree with the FemTechNet whitepaper, which states that "We seek to activate a learning process that recognizes and extends across global and cultural contexts. While the use of the world-wide-web and internet infrastructures enables communication among people at great geographic distances, it also strains the capacity for respect and the appreciation of the nuances of diverse backgrounds which increases the intensity of the work that must be done by teachers and organizers of the learning process."³ In our paper we will discuss how the structure of the project is shaped by such concerns. Second, our project focuses on the recovery of cultural objects that have not been included in digital collections. Recent work, particularly self defined postcolonial digital humanities projects, has pointed to the lack of attention that digital humanities pays to historical and literary production by marginalized peoples. Our efforts are focused on the newspapers, photographs, historical accounts and other archival artifacts that discuss the racial violence, tensions and other aggressions (micro and macro) in our localized Texas environment. By digitizing a related collection of materials we will diversify the digital canon related to conceptions of Texas and the universities. Through our digitization project students will be theoretically and methodologically immersed in practices of research and archival decision making critical to a broader more diverse digital canon.

This project marks an expansion of previously developed techniques. Our work grows out of individual recovery projects that have been tested by both Taylor and Earhart. Taylor has worked with her undergraduate students to collect the oral histories of women who have had a thirty year or longer relationship to Prairie View A&M University as faculty, staff and/or alum/student in the development of the Prairie View Women Oral History Project. Earhart has worked with her undergraduate and graduate students to recover the Alex Haley Malcolm X papers (*Scholarly Editing* 2014) and selected Black Radicals Papers which include materials from the Black Panther

movement, Angela Davis' prison stay, and various racial reform protests in the 1960s.

We have selected a well-known technology tool as a technological interface for collaboration. Not only is Omeka an excellent entry technology because of its low cost, simple interface, and large developer and user community, but we argue that Omeka is useful as teaching tool due to the emphasis on Dublin Core metadata which forces students to engage with the ways that a controlled vocabulary has been culturally shaped and where such vocabulary is at odds with specific cultural norms engaged in the black resistance documents.

Each class is working with collections held at their respective libraries to find the artifacts for the project then collaborate to curate a cohesive digital space which tells the story of White Violence and Black Resistance in Texas. Using Omeka as our bridge, Earhart and Taylor have modeled student research across the two universities, emphasizing individual archival collection and collaborative moments of interaction between the classes. This paper will discuss the challenges with working across disparate universities where tradition and necessity have created the space for PVAMU to thrive, its one hundred and thirty plus year history itself a statement of resistance. This project expands the canon to understand the intersectional ways in which both universities actively and passively engage in moments of violence and resistance. Each class has focused on a particular historical event that occurred in the local area. Earhart's class is investigating what we believe to be the largest "race riot" in Texas. The reconstruction era conflict occurred in Millican, Texas, a town located 15 miles from the Texas A&M University campus. During the first KKK rally in Millican, in 1868, armed freedmen fired on the rally, driving the Klan out of town. After the rally, George Brooks, a local preacher, began a black militia. Several confrontations occurred including a march on Bryan by a large group of armed blacks, which ended in an armed conflict and lynching of Brooks and from 5 to 100 African-American men, women and children. Reports of the conflict were recorded in newspapers around the world. In addition, the Freedmen's Bureau has records from those sent to investigate the incident. This sketchy information promises to prove fertile for investigation of what is a pivotal event in Texas and African-American history. Taylor will be investigating the ways in which this and other events serve as part of the cultural narrative historical legacy of students, faculty and staff of Prairie View's earliest members. Students are working with university and community archives, historical newspapers, local church archives, interviews, local Blues lyrics and local literary narratives. During the class, the students locate materials, digitize, transcribe, conduct historical research on the document and related people and places, and enter the materials and metadata into Omeka. We will share materials developed for our classrooms that guide students through Omeka. Of particular interest will be the materials that we have created to explain and facilitate the use of Dublin Core metadata. In addition, we will highlight ways that we have focused interactivity between the separate institutions and continuing challenges of such work. Our pedagogical approach not only helps students to understand archive to digitization projects, but emphasizes critical engagement in technological decisions.

We will highlight what our project has taught us about developing a more comprehensive, inclusive and effective digital pedagogy and discuss future steps in our collaborative project. Once a substantial set of materials are collected, future classes will construct timelines and maps of the occurrences. By treating the materials as data, using a controlled set of metadata and search techniques, we will be recovering important information that might interact with broader data sets of materials nationally and internationally, ensuring that these important events are not erased from history. Alan Liu has challenged digital humanists "To be an equal partner—rather than, again, just a servant—at the table," by finding "ways to show that thinking critically about metadata, for instance, scales into thinking critically about the power, finance, and other governance protocols of the world."⁴ Our project provides a replicable model that we believe will empower other teachers and students to engage with digital humanities.

References

1. Earhart, A. (2012). *Can Information be Unfettered? Race and the New Digital Humanities Canon*. In Gold, M. K. (ed), Debates in the Digital Humanities. Minneapolis: University of Minnesota Press. 309-318.
2. FemTechNet.
3. FemTechNet. (2013). FemTechNetWhitepaper. femtechnet.newschool.edu/femtechnet-whitepaper
4. Liu, A. (2012). *Where Is Cultural Criticism in the Digital Humanities?* In Gold, M. (ed), Debates in the Digital Humanities. Minneapolis, MN: University of Minnesota Press, 490-509.

Potential Criticism in the Digital Humanities

Edwards, Richard

redwards7@bsu.edu
Ball State University

In this paper, I argue that the concept of potential criticism addresses many key challenges and issues involving cultural empowerment processes in the digital humanities. With the proliferation of digital humanities projects that engage with the study and analysis of large sets of digital texts, potential criticism proposes an analytical methodology that has broad applicability across different types of texts, media, and disciplines. At its core, potential criticism is an idea derived from the work of the Oulipo and recombinatory poetics. The Oulipo is an acronym for the Ouvroir de literature potentielle, which roughly translates into English as "Workshop of potential literature." The writers and mathematicians of the Oulipo focused on creating new works through the use of constrained writing techniques. Well-known members of the Oulipo include Georges Perec, Italo Calvino, and Raymond Queneau.

But constrained and algorithmic techniques can not only be used to generate new literary works, such methods also can be used to analyze existing texts, as evidenced by the work of the Oulipo's Harry Mathews and his Mathews's Algorithm. Mathews advances the idea that pre-determined and mathematical constraints can be used as a way of both recombining and analyzing existing texts. As Mathews states as the beginning of his essay explaining his algorithm: "Potential reading has the charm of making manifest the duplicity of texts, be they oulipian or not." (trans. Shannon Clute). This work has important consequence for cultural empowerment processes in the digital humanities. As we seek to analyze ever-larger bodies of digital texts through our digital humanities projects, we need new methodologies that can extract meaningful data sets for further analysis and discovery.

In our 2011 book, *The Maltese Touch of Evil: Film Noir and Potential Criticism* (Dartmouth College Press), Shannon Clute and I laid out how the methodology of potential criticism could be used to analyze moving image texts (specifically in our case, films noir). Our method has important implications for the digital humanities, which I will expand upon in this paper. First, we focused on a moving image archive to show how constrained analytical techniques can move beyond the analysis of alphabetic and/or literary texts. Potential criticism can be used on any digital text or medium. Second, echoing the work of Stephen Ramsay who deploys a similar methodology in his book *Reading Machines: Toward an Algorithmic Criticism*, the results of potential criticism are not an end in and of themselves. These new data sets, revealed through the application of mathematical or algorithmic means, produce new and generative starting points for further investigation and discovery. Furthermore, the resultant data sets are not merely a random remix of a body of texts, they are constrained data sets that reveal potentially new information about the overall corpora itself.

This paper argues for the broader applicability of the concept of potential criticism in the digital humanities. I will highlight

two practical demonstrations of potential criticism as a working methodology. First, we used the concept of potential criticism in the Film Annotator's Workbench Project from Indiana University (IU) in conjunction with IU programmer Will Cowan. Cowan's work on this project was funded by the NEH. In this project, which began in 2010, we used pre-existing digital humanities tools (in this case, the Film Annotator's Workbench and Omeka) to demonstrate how the potential criticism methodology can analyze and annotate a large number of films noir. Second, we similarly used the concept of potential criticism as the basis for our new investigations around films noir, which was published in 2011.

The goal of this paper is to share our findings and disseminate our approach to potential and algorithmic criticism. I will contribute my first hand observations from our multi-year investigations into how potential or algorithmic criticism can be used to analyze any kind of digital text and media, especially our work around film and video analysis and annotation. Moreover, I will discuss how potential criticism as a working method can empower certain types of scholarly communities around shared corpora of texts. Both the Film Annotator's Workbench Project and our Maltese Touch of Evil Project operate on the open web through freely available tools to publicly disseminate constrained data sets for other scholars to explore and discover.

In keeping with the theme of this year's DH conference, potential criticism as a practice intentionally encourages scholars to work across disciplinary boundaries to share their insights and ideas in a workshop fashion. In many ways, potential criticism needs to be an open-ended practice that does not presuppose the primacy of any particular hermeneutic method, but rather encourages a variety of approaches in order to reveal the potential information that exists within large bodies of texts. In closing, potential criticism can operate as a method for rapidly testing out new hypotheses that will advance our cultural understanding of large corpora of texts and media. This is one of the benefits, but also one of the key challenges, of potential criticism. Each algorithmically derived data set is novel investigation into a body of texts. Therefore, potential criticism generates its own kind of scholarly archive that requires consideration and development. It is imperative among digital humanities scholars, researchers, and programmers to continue to explore and build new ways of making newly generated data sets readily available to other scholars through the open web.

Sequence, Tree and Graph at the Tip of Your Java Classes

Eide, Øyvind

oe@oeide.no
Universität Passau

1. Introduction

How to represent texts in computer systems has always been an important topic in Digital Humanities. Tree based formalisms such as SGML¹ and XML (URL: <http://www.w3.org/XML/> (checked 2013-10-26)) are useful for many purposes, but problems related to their hierarchical structures are inherent. Various solutions have been presented over the years, from questioning the existence of overlap in textual material² through various workaround for overlapping structures³ (chapter 20) to the abolition of nesting formalisms, as in the example of MECS⁴.

In this paper I will focus on three different design principles for text representation systems, namely, linear, hierarchical and graph based. These labels represent concepts similar to Cayless' data types text as stream, text as tree, and text as graph⁵. Examples of text modelling tool types focusing on

each of these can be found in table 1. In the following I will analyse the relationship between the left and the right sides of the table based on experiences from the development and use of a text modelling tool called GeoModelText (URL: <http://sourceforge.net/projects/geomodel/> (checked 2013-10-26)). Description of the development process can be found at the resource page for my PhD project (URL: <http://www.oeide.no/dg/dp/> (checked 2013-11-01)). A use case is described in⁶.

Type	Text representation system	Example of modelling tool type
1	Linear	Plain text
2	Hierarchical	XML encoding
3	Graph based	RDF encoding

Table 1 caption: Computer based text representation systems and modelling tools.

2. Previous work

In the last 10–15 years, most practical work in text encoding have lived with the nesting structure of SGML and XML. Alternative formalisms, such as MECS, has mostly yielded to the dominant tree based structure. There has, however, been an undercurrent of experiments and theoretical research into other types of solutions. One important example originally initiated in 2002 is LMNL (URL: <http://www.lmnl-markup.com/> (checked 2013-10-26)), with its range based annotation⁷. A recent attempt to examine text markup under the microscope is *pure transcriptional markup*⁸. Both these examples show ongoing attempts to get to the core of markup practice and were important in the development of this paper.

In my own research into semiotic differences between texts and maps I developed a system for computer assisted conceptual text modelling called GeoModelText where all three types from table 1 have the same status. GeoModelText is currently tailor suited to my work. One aim of this paper is to investigate into its usability for other types of research.

Systems for visualising graph networks based on and connected to texts are easily available. What is lacking are systems for manipulating the graph structure as part of interactive text modelling, that is, including the graph structure in the internal editing system. To the degree graph based editing is included in XML editors it tends to be in an indirect sense, e.g., through manipulation of attribute values for ID-IDREF links.

3. GeoModelText

GeoModelText is implemented as a Java application. The text to be analysed is imported from a TEI P5 document. The import results in a DOM structure representing the XML tree of the TEI source. Within the DOM structure the text itself is found in PCDATA segments. During analysis, these segments are used to reconstruct parts of the text as a linear structure; in this view the XML hierarchy is hidden from the user.

An important function of a tree structure is inheritance, where aspects of a parent is inherited by its children. This is used in the system to migrate responsibility to statements: the person responsible for a paragraph is also responsible for each sentence in the paragraph. One example is relationships between places claimed in the text. The claims are connected to the person responsible for the encapsulating paragraph. The places are referred to by names. In order to build up networks of related places, co-references between occurrences of names must be taken into consideration. For an introduction to co-reference see⁹, and¹⁰ for the use of co-reference networks. Thus, in order to establish this graph structure of related places, we need to read the linear text (type 1 in table 1), use inheritance (type 2), and then build up a network of related places and place names (type 3). This network connects strings which can be found at any level of the document structure. The

co-reference networks change the document structure as a whole into a graph, as indicated in figure 1.

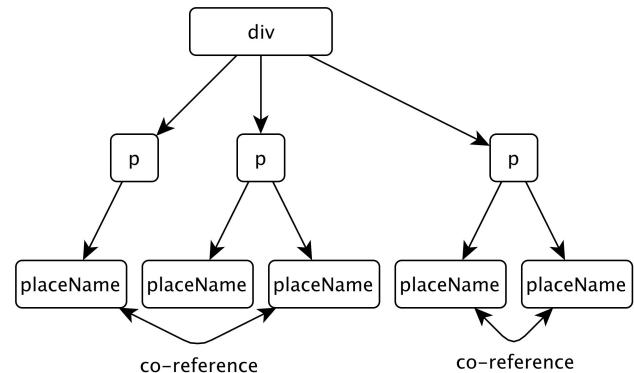


Fig. 1: The XML tree turned into a graph.

In the graph data structure, each node has a chain of links back to the place in the tree, and thus in the text, on which it is based. In this perspective, type 3 from table 1 is the dominant, but with links to the other types. In the distinction in¹¹ (p. 75) between hypertextual editions and editions in database format, GeoModelText is primarily a database system, but with aspects of a hypertext system.

4. The hierarchy of structures

Everything we have seen in the example above is well known from XML tools, they are commonly used to establish such data structures. However, in the typical XML tool, the status of the three types are presented differently. To take one example: when opening an XML file in the Oxygen XML editor (URL: <http://www.oxygenxml.com/> (checked 2013-10-27)) the tree structure is presented at the left of the Oxygen window as a number of expandable folders. This indicates that this structure is the base structure of the document. In the centre window we see the text as a sequence of tokens, with the XML elements shown in different ways or fully hidden at user discretion. Graph structures, e.g., a co-reference network, is only shown indirectly as attribute values.

Thus, type 1 and 2 from table 1 are highlighted at the expense of type 3. In GeoModelText, on the other hand, separate windows give access to data in any of the three types. I do not claim that the graphical user interface of GeoModelText is better than the one in Oxygen; it is not. Rather, I want to make the point that the freedom of the object oriented programmer is not yet offered to the user of markup tools. What a user actually do with markup tools can be expressed in three different layers:

1. Create a tree model of the text
2. Formalise that model in TEI/XML
3. Reifycate the model into the editing tool

With these layers reified into the tool it is more difficult to get rid of, or even see, the straightjacket of the tree structure. While XML is useful for many purposes and is used both as input format and as storage format for GeoModelText, there is still a tendency to under-expose the non-hierarchical structure of the text. This tendency is there even if a number of techniques are developed to overcome the problems created by the hierarchical nature of XML.

A radical solution to these problem would be to leave XML, or, more generally, avoid hierarchical markup systems based on context free grammars all together. This was the solution of MECS. However, this solution generated its own set of problems. In the development of LMNL, the choice was rather to use XML and the XML tools for what it is good for, and only leave the hierarchical structure when one needs something else. This is also used as a design principle in GeoModelText. I use XML, both in the form of linearised files and DOM structures, whenever it is useful. But by operating in a programming environment (which is also where Piez finds himself in his work with the LMNL toolchain Luminescent (URL: <https://github.com/>

wendellpiez/Luminescent/ (checked 2013-11-01)), although the language is different), I can leave the XML/DOM structure whenever I need to, e.g., to establish tools for capturing and visualising co-references.

5. The way forward

Sometimes, in the frustration of the hierarchical straightjacket, it is tempting to leave XML as a whole behind. Keeping XML as a part of the information system, adding non-hierarchical modules as needed, is a better way forward for text encoding in general and for TEI specifically. XML is only a straightjacket in certain situations. Luminescent and GeoModelText are two examples of tools pointing towards a future where we can live with the limitations of XML because we are free to leave it whenever we need to. Craig outlines a complementary strategy for improvements of TEI¹².

One central question remains, however. Is it possible to avoid the straightjacket as a user of a premade system, or is it something fundamental in my role as a programmer who gives me the freedom to create input mechanisms as well as visualisations based on all three types from table 1? Can we make tools which give users this freedom? When we create a tree model of a text we are still aware of the choices made, often out of convenience. When using an XML editing tool these choices have been implemented in a structure with a real material existence. The awareness of the straightjacket may be lost in the process from abstract model to editing tool.

References

1. Goldfarb, C. F. and Y. Rubinsky (1990). *The SGML handbook*. Oxford: Clarendon Press.
2. Renear, A., E. Myonas, and D. Durand (1996). *Refining our Notion of What Text Really Is: The Problem of Overlapping Hierarchies*. In S. M. Hockey and N. M. Ide (Eds.), Selected papers from the ALLC/ACH Conference, Christ Church, Oxford, April 1992, pp. 263–280. Oxford: Clarendon Press.
3. TEI Consortium (2013). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. [2.5.0]. [July 26 2013]. [S.n.]: TEI Consortium.
4. Sperberg-McQueen, C. and C. Huitfeldt (1999). *Concurrent document hierarchies in MECS and SGML*. Literary and Linguistic Computing 14(1), 29–42.
5. Hugh A. Cayless, Rebooting TEI Pointers. In *Journal of the Text Encoding Initiative*, 6(2013). URL: jtei.revues.org/907 (checked 2014-03-07) ; DOI: 10.4000/jtei.907
6. Eide, Ø. (2013). *Why Maps Are Silent When Texts Can Speak*. Detecting Media Differences through Conceptual Modelling. In M. F. Buchroithner, N. Prechtel, D. Burghardt, K. Pippig, and B. Schröter (Eds.), Proceedings from From Pole to Pole. The 26th International Cartographic Conference 2013, Dresden.
7. Piez, W. (2013). *Markup Beyond XML*. In Proceedings from Digital Humanities July 16–19 2013, University of Nebraska–Lincoln, USA, pp. 343–345. Center for Digital Research in the Humanities.
8. Caton, P. (2013). *Pure Transcriptional Markup*. In Proceedings from Digital Humanities July 16–19 2013, University of Nebraska–Lincoln, USA, pp. 140–142. Center for Digital Research in the Humanities.
9. Eide, Ø. (2009). *Co-Reference: A New Method to Solve Old Problems*. In Proceedings from Digital Humanities June 22–25 2009, University of Maryland, USA, pp. 101–103. Maryland Institute for Technology in the Humanities.
10. Meghini, C., M. Doerr, and N. Spyros (2009). *Managing Co-reference Knowledge for Data Integration*. In Y. Kiyoki, T. Tokuda, H. Jaakkola, X. Chen, and N. Yoshida (Eds.), Information Modelling and Knowledge Bases XX. Amsterdam, Washington, D.C.: IOS Press.
11. Buzzetti, D. (2002). *Digital Representation and the Text Model*. New Literary History 33(1), 61–88.
12. Hugh A. Cayless, Rebooting TEI Pointers. In *Journal of the Text Encoding Initiative*, 6(2013). URL: jtei.revues.org/907 (checked 2014-03-07) ; DOI: 10.4000/jtei.907

Exploratory Thematic Analysis for Historical Newspaper Archives

Eisenstein, Jacob

Georgia Institute of Technology, United States of America

Sun, Iris

Georgia Institute of Technology, United States of America

Klein, Lauren F.

lauren.klein@lmc.gatech.edu

Georgia Institute of Technology, United States of America

Introduction

On July 19th, 1848, 300 concerned United States citizens gathered in Seneca Falls, New York, for the women's rights convention that would culminate in the signing of the Declaration of Rights and Sentiments, the first major document (in the US) to call for women's right to vote. In *The North Star*, Frederick Douglass, the former slave turned abolitionist, extolled the event as a "grand movement for attaining the civil, social, political, and religious rights of women" (1848). In the *Oneida Whig*, the same event was ridiculed as the "most shocking and unnatural event ever recorded in the history of womanity" (1848). As demonstrated by these contradictory accounts, published opinions varied greatly -- about the women's rights movement in the nineteenth-century United States, and about current events generally conceived. Large-scale digitization projects have increasingly enabled humanities scholars to search newspapers, such as those just cited, for significant words and phrases. But exploring more open-ended questions such as, "How did the discourse surrounding women's rights in the United States change in the wake of the 1848 Seneca Falls Convention?" or "Did the women's rights movement borrow language from the nation's contemporaneous anti-slavery campaign?" remains a challenge. Synthesizing current research on exploratory data analysis with techniques from the fields of computational linguistics and data visualization, we propose a new set of methods to assist humanities scholars in computationally-assisted exploratory research.

Background and Overview

Exploratory data analysis (EDA) has played a fundamental role in quantitative research since at least the 1970s (Tukey 1977). In comparison to formal hypothesis testing, exploratory data analysis is more open-ended, and is meant to help the researcher develop a general sense of the properties of the dataset before embarking on more specific inquiries (Russell, Stefk, Pirolli, and Card, 1993). EDA typically combines visualizations such as scatterplots and histograms with lightweight quantitative analysis, serving to check basic assumptions, reveal errors in the data-processing pipeline, identify relationships between variables, and suggest preliminary models. More recently, Andrew Gelman (2004) has argued that EDA should be interwoven with formal statistical modeling, facilitating an iterative design process driven by experimenter insight.

The questions about women's rights, posed above, suggest the potential of EDA for humanities research-- a possibility also noted by Muralidharan and Hearst (2012). That team employs automatic syntactic analysis to identify and visualize recurring grammatical patterns, which, when combined with document metadata, reveals insights at the sentence level. By contrast, we combine metadata with techniques such as topic modeling in order to reveal insights at the document level. Inspired by the increasing use of topic models to make literary and cultural arguments (Underwood 2012, Rhody 2012, and Jockers 2013), we ask how the *exploratory thematic analysis* of documents might be incorporated into the initial phase of humanities research.

Our approach encompasses both traditional topic models and innovative visualizations, as well as alternative computational techniques targeted at the questions that topic modeling raises but leaves unanswered. By designing new visualizations and text-mining algorithms within the context of a specific, humanities-driven research effort, we hope to prototype a new mode of multi-disciplinary scholarship that will facilitate the iterative research methodology advocated by Gelman (2004). Specifically, we aim to facilitate the thematic exploration of document archives as a precursor to more informed keyword searching, more sustained close reading, and more systematic evidence gathering.

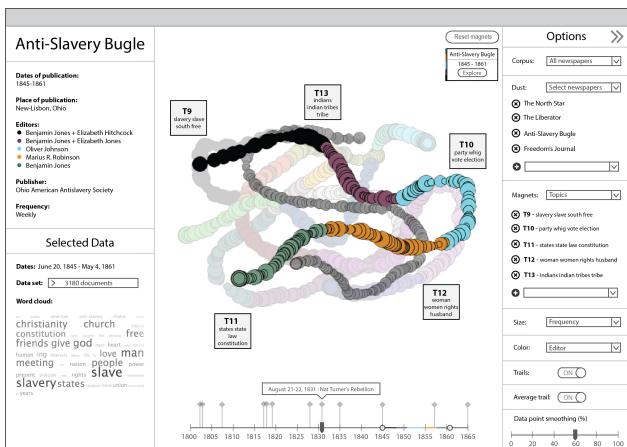
User Scenario and Interface Prototypes

Our focus is on a set of abolitionist newspapers from the nineteenth-century United States, in which antislavery advocates mounted moral, social, and political arguments in favor of emancipation. These newspapers present a particularly compelling dataset for thematic analysis, as similar ideas were purportedly framed differently by (and for) women and men (Dudden 2011). Here, we focus on one newspaper, *The Anti-Slavery Bugle*, published in New Lisbon, Ohio, between 1845 and 1861. Significantly, it was the source of much reprinting (Golden 2013), and underwent several shifts in editorial control.

Standard LDA topic analysis (MALLET; McCallum 2002) with 100 topics and standard parametrization reveals a number of topics that might intrigue a scholar in the initial phases of research, including:

- T40: states state law constitution the government power united laws congress **rights** people con ohio tion act union question property
 - T56: indians indian tribes tribe chiefs frontier dian treaties tiger hawk antelope annuity fiscal lllack hyenas tigers dians avalanche savages
 - T59: woman women **rightshusband** wife sex sho marriage property married mrs female legal sphere equality estate social duties sexes

Topic 59 (T59) suggests that the *Bugle* may offer insight into the relationship between the antislavery movement and the nascent drive for women's rights. The accompanying metadata reveals that the newspaper was co-edited by a woman between 1845 and 1849; however, this topic peaked in the late 1850s (a time when the women's rights movement was ascendant). At this point, we reach the limit of what can be learned from a topic model alone. We cannot easily answer, for instance, how the treatment of this topic in the *Bugle* might have differed from that of other newspapers; whether the editors of the *Bugle* were early advocates for the women's rights movement; or whether the peak in the late 1850s followed a national trend.

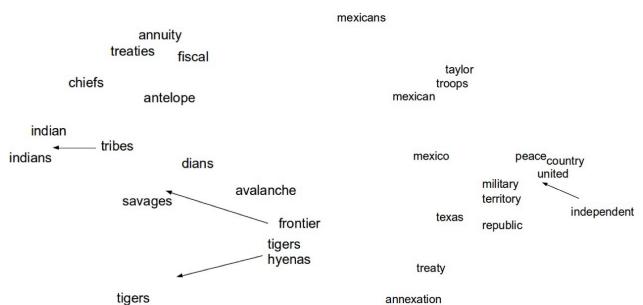


The above screenshot documents a prototype interface for the visualization and analysis of topic models that can begin to answer these questions. We apply a *dust-and-magnet* visualization (Yi et al., 2005), in which user-selected topics exert a magnetic “force” on individual issues of the

newspaper (represented as “dust”). The temporal trajectories of several newspapers are shown as “dust trails” in the visual space, with colors indicating the terms of different editorial teams, and with the *Bugle* highlighted so as to facilitate comparison with contemporaneous newspapers.

Next, we address the topics themselves. In such models, topics are defined by sets of words, with the assumption that each word has a single meaning across all usage contexts. However, much humanities scholarship entails a sensitivity to shifting meanings and uses. A scholar may wonder, for instance, how women’s “rights” (as indicated by the keyword in T59) were described in relation to the legal “rights” featured in T40. She may ask if the rhetoric of one borrowed from the other, or if the use of the word “rights” changed when it was employed to describe women’s vs. legal rights. Again, the scholar seeks to know more than what can be inferred by LDA alone. We propose to link LDA’s high-level thematic analysis with visualizations that drill down to the level of individual examples. Building on the traditional keyword-in-context (KWIC) models, we are developing a computational algorithm for selecting contexts that are both strongly associated with each topic of interest (for example, the contexts for “rights” in T40 and T59), while simultaneously revealing the full range of thematic possibilities within each topic.

While the range of connotations of individual words in a topic presents one kind of interpretive challenge, the topics themselves can at times present another: when a topic includes words associated with seemingly divergent themes. In T56, the scholar might observe a (seemingly) obvious connection, for the nineteenth-century, between words that describe Native Americans and those that describe nature. However, unlike the words “antelope” or “hawk,” the words “tiger” and “hyena,” also included in the topic, do not describe animals indigenous to North America. Does an explanation lie in a figurative vocabulary for describing native peoples? Or is this collection of words merely an accident of statistical analysis, a result of being built on a randomized algorithm?



To address this question, we propose a spatial visualization using multidimensional scaling (Cox and Cox, 2010) to position the keywords for each topic according to their contextual similarity. As shown in the figure above-left, the terms “indian”, “indians”, and “tribes” are located apart from “hyena”, “tiger”, and “tigers”, which are themselves closely associated. The spatial layout suggests a relatively weak connection between these terms. For comparison, we also include the spatial visualization for a topic relating to the Mexican-American War, above-right, in which terms related to the conduct of the war (“Taylor”, “troops”) are spatially distinguished from those related to its outcome (“treaty”, “annexation”).

Conclusion and Next Steps

The goal of our ongoing work on exploratory thematic analysis is to provide a comprehensive set of algorithms and visualizations for understanding newspaper archives. Topic modeling is an important first step, but if we are to move beyond suggestive word lists in order to contribute to humanities scholarship, topic models must be linked to relevant metadata and concrete examples. Moreover, scholars must be provided with new visual modes that illuminate the substructures within the generalized themes that the topic model produces. Such techniques can reveal new insights about the transmission and

circulation of ideas among social and political coalitions, and how the framing of these ideas relates to authors' genders. By linking technical innovation with real humanistic inquiry, we hope to produce algorithms and visualizations that will meet the needs of substantive humanities research.

References

- "*Bolting Among the Ladies.*" The Oneida Whig, August 1, 1848.
- Cox, Trevor F., and Michael AA Cox** (2010). *Multidimensional scaling*. CRC Press.
- Douglass, Frederick** (1848). "*The Rights of Women.*" The North Star, July 28.
- Dudden, F.** (2011) *Fighting Chance: The Struggle Over Woman Suffrage and Black Suffrage in America*. Oxford UP.
- Jockers, M.** (2013) *Macroanalysis: Digital Methods and Literary History*. U Illinois P.
- Kwok, James T. and Ryan P. Adams.** (2012). "Priors for Diversity in Generative Latent Variable Models." Advances in Neural Information Processing Systems.
- Rhody, L.** (2012) "Some Assembly Required: Understanding and Interpreting Topics in LDA Models of Figurative Language." Lisa @ Work.
- McCallum, Andrew Kachites** (2002). "MALLET: A Machine Learning for Language Toolkit."
- Muralidharan, Aditi. and Hearst, Marti A.** (2012), *Supporting Exploratory Text Analysis in Literature Study*, Literary and Linguistic Computing, 27 (4), Dec 24.
- Russell, Daniel M., Mark J. Stefk, Peter Pirolli, and Stuart K. Card.** (1993) "The cost structure of sensemaking." In Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems, pp. 269-276. ACM.
- Tukey, John Wilder** (1977). *Exploratory Data Analysis*. Addison-Wesley.
- Underwood, Ted** (2011). "*The Differentiation of Literary and Nonliterary Diction, 1700-1900.*" The Stone and the Shell.
- Yi, Ji Soo, Rachel Melton, John Stasko, and Julie A. Jacko** (2005). "Dust & magnet: multivariate information visualization using a magnet metaphor." *Information Visualization* 4, no. 4: 239-256.

Literary Canon and Digital Bibliographies: The Case of the United States

Ferrer, Carolina

ferrer.carolina@uqam.ca
Université du Québec à Montréal

In this research, I propose an alternative technique to the traditional method of constitution of the literary canon. Instead of basing the determination of the canon on different values, I scrutinize the *Modern Language Association International Bibliography* database in order to determine the most cited authors and literary works. Specifically, I study the literature of the United States of America. Thus, through the process of data mining, I obtain a sample of over 290,000 references that allows us to observe the chronological evolution and the linguistic distribution of the critical bibliography about USA literature. This quantitative technique yields a corpus of more than 100 titles and 100 writers that are cited more than 100 times in the database. Consequently, this bibliography is not the result of subjective selection criteria, but is based on the law of large numbers. Furthermore, this study shows that the quantitative analysis of bibliographic databases is an effective way to bring new light to the field of literary studies.

μServices and The Riddle of Literary Quality

Filarski, Gertjan

gertjan.filarski@huygens.knaw.nl
Huygens ING, Netherlands, The

de Jong, Hayco

hayco.de.jong@huygens.knaw.nl
Huygens ING, Netherlands, The

van Dalen-Oskam, Karina

karina.van.dalen@huygens.knaw.nl
Huygens ING, Netherlands, The

Introduction

The Riddle of Literary Quality is a project funded by the Computational Humanities Program of the Royal Netherlands Academy of Arts and Sciences (KNAW). It runs at Huygens ING in partnership with the Institute for Logic, Language and Computation of the University of Amsterdam, and the Frysk Akademy in Leeuwarden. The aim of the project is to develop a method and the necessary software to analyze low-level and high-level formal features in a corpus of modern Dutch long fiction, to find out whether formal features in the texts play a role in the reception and evaluation of the text by the readers. Can we get more insight into the responses of readers to, for instance, texts with on average longer versus shorter sentences, or using a larger vocabulary, or on average showing a more complex syntactical structure (cf. Jautze et al.)? Is there a difference between those texts that readers consider to be highly literary and those that are experienced as more lowbrow? Can we distinguish texts found good or bad by readers based on formal features in these texts? And how do the opinions of readers correlate with the kind of reader they are?

The project thus aims to correlate formal features with readers' opinions and readers' roles. The analysis of the formal features is done through a chain of μServices that we will deal with in the second part of this paper. The first part is addressed to the analysis of readers' opinions and readers' roles.

Survey

To gather information about readers and their responses we set up a large online survey in which we asked respondents some personal information (age, gender, postal code, level of education) and sixteen questions to find out what kind of reader they predominantly are: autonomous or 'distanced': reading for aesthetical pleasure; or heteronomous 'identifying': reading for fun, to discover other cultures or places, or to identify with the main characters. We based our distinction and our questions on work done by Von Heydebrand & Winko on sociological aspects of (literary) reading. Next to that, we presented a list of 400 recent novels, Dutch originals or translations into Dutch, and asked them to mark the ones they read. A selection of these novels was presented to them, with the question to evaluate these works on two scales: from 'not so very literary' to 'highly literary' and from 'bad' to 'good'. The survey ran for six months, and received almost 14000 respondents. Analysis of the results has just started.

The results of the survey will be correlated with the results of the measurements of formal features. We would like to describe the technical set-up we have devised to enable the scholars to analyze the texts in the corpus in a way that is trustworthy and sustainable, using μServices written in Java that can also be used by others to repeat and to verify our analyses.

μServices

Research infrastructure for the Riddle of Literary Quality is designed with three goals in mind: research results must be reproducible; analytical tools must be reusable; the entire

workflow must be maintainable and reliable. We aim to provide a toolset that allows for a verifiable system that will focus the discussion on the selected methodology - the procedures and algorithms. This also means that we will make the code behind each µService open source.

To accomplish this goal the digital humanities engineering group at Huygens ING based the research infrastructure both on the results of COST Action IS0704: *An Interoperable Supranational Infrastructure for Digital Editions* (Interedition) – of which the institute was grant-holder – and the work of Joris van Zundert – chair of the Action. Van Zundert (Huygens ING) specified as an objective of the COST Action the development of lightweight and distributed interoperability solutions. These solutions were implemented through webservices. The CollateX algorithm of Ronald Haentjens Dekker (Huygens ING) and Gregor Middell (University of Würzburg) was among the first and most successful of a series of compact analytical demonstrators called µServices.

The Riddle of Literary Quality does not aim to build a workflow management system. Such a top-down standardization methodology is left to large infrastructural programs like CLARIN, DARIAH or the Dutch Nederlab project. Instead we continue Interedition's grassroots approach and leave (computational) researchers and PhD students free to experiment with high-level and low-level analytical algorithms in languages that range from Python to Java. These algorithms may or may not grow out to be part of the Riddle's µService infrastructure and those that are deemed useful are eventually hosted at the institute's servers.

The current services fall in three distinct categories: data import and preparation; analysis and visualization and export. In the first group we offer e.g. a series of tools that convert documents to specified standards (such as ePub/PDF to TEI) and set the data in the correct character encoding (such as a conversion from Windows-1252, ISO8859 to UTF-8). To prepare the data for further analysis we have converted parsers like the Dutch Ucto: Unicode Tokenizer (Radboud University of Nijmegen/University of Tilburg) to a µService. The output data of services in this group is a standardized json format that can be read by the analytical services in the second category. Experiments in The Riddle currently focus on this analysis group. µServices in the third category perform output operations. Some create visualizations while others export the data to external environments for further analysis. For stylometric research e.g. we created a µService to export data from The Riddle to R and integrate it with the Stylo() package created at the Universities of Krakow (Macej Eder/Jan Rybicki) and Antwerp (Mike Kestemont).

The entire suite of µServices will remain available for persistent access and may be used in alternate workflows or by external third-party software. Thus the suite does not only allow reproduction of the results of The Riddle but will also support entirely new and original research.

Sample Workflow

As an example of a µServices-driven workflow we present one possible use of gathering statistical data from a corpus of eBooks. First, each eBook is sent to a service that prepares it for analysis by converting the book into a structured TEI document. Character-encoding issues are resolved by a second µService, resulting in a normalized, platform-independent UTF-8 version of the TEI-document. Subsequently, a third service offers extraction operations on the structural level of the file. This service is used to extract all relevant paragraphs. These paragraphs are split into sentences and words by one of a family of (TEI agnostic) tokenizers, such as the Ucto-µService. Statistical analysis of these tokens is possible by sending the resulting list of tokens to the exporter µService, which transforms the extracted tokens into a format suitable for use in R.

Conclusion

To make sure that we are able to answer the main questions of The Riddle of Literary Quality – whether there are any correlations between readers' opinions about certain novels, readers' predominant reading role, and the values for a list of formal low-level and high-level features of the novels – we have chosen to develop a set of µServices that deal with single aspects of the needed analysis. By making these µServices available to other scholars we enable them to repeat and verify our research results. We provide users with tools that can be used to answer different questions than we have in The Riddle, thereby making the tools also useful in a wider sense for new original research. We hope our approach invites others to contribute µServices for further textual humanities research.

References

- CollateX*. collatex.net
- Interedition*. www.interedition.eu
- >*Riddle of Literary Quality*. literaryquality.huygens.knaw.nl
- Ucto*. ilk.uvt.nl/ucto
- Eder, M., Kestemont, M. & Rybicki, J.**, (2013). *Stylometry with R: a suite of tools*.Digital Humanities 2013: Conference Abstracts. Lincoln: University of Nebraska-Lincoln, pp. 487-89. dh2013.unl.edu/abstracts/ab-136.html
- Heydebrand, R. von and Winko, S.**, (1996), *Einfuehrung in die Wertung von Literatur. Systematik – Geschichte – Legitimation*. Paderborn etc.: Ferdinand Schoeningh, 1996
- Jautze, K., Koolen, C., Cranenburgh, A. van and Jong, H. de**, (2013). *From high heels to weed attics: a syntactic investigation of chick lit and literature*.Proceedings of the Workshop on Computational Linguistics for Literature 2013. aclweb.org/anthology/W/W13/W13-1410.pdf
- Zundert, J. van, Middell, G., Hulle, D. Van, Haentjens Dekker, R., et al.**, (2011). *Interedition: Principles, Practice and Products of an Open Collaborative Development Model for Digital Scholarly Editions*. Digital Humanities 2011: Conference Abstracts. Stanford: Stanford University. dh2011abstracts.stanford.edu/xtf/view?docId=tei/ab-227.xml

From Markup to Analysis: Culture Claims and Code in the Digital Archive

Flanders, Julia

j.flanders@neu.edu
Northeastern University

Dillon, Elizabeth Maddock

e.dillon@neu.edu
Northeastern University

Following the lead offered in the "Text Encoding Meets Text Analysis" panel session at DH2012 in Hamburg,¹ and also in the 2013 debate between Matthew Jockers and Julia Flanders on "A Matter of Scale",² the research value of basic modeling for text analysis seems comparatively clear, even if it is as yet unrealized in practice. However, the payoff from the more complex forms of modeling that are evident in typical TEI-encoded thematic research collections has been less clearly demonstrated. The usefulness of such modeling is evident insofar as it supports the publication of these collections: information about document structures is used to produce display formatting and navigation, and information about content features such as named entities and genre is commonly used in searching. But the larger claims made for digital resources—that they "allow new questions to be asked"³—require us to distinguish between these kinds of functional value and what we might call "real research value." The real research value of the modeling is reflected in the ability of the data to support complex inferences, at scale, that materially contribute to humanities research arguments. Huitfeldt et al. in "Meaning and Interpretation of Markup"⁴ offers a foundational

demonstration of how markup "licenses inferences" in formal terms. But projects developing TEI-encoded archival collections have not yet articulated in detail how that markup might lead researchers from comparatively simple inferences ("this paragraph contains a reference to such-and-such a person") to more complex ones ("if the name of a person classified in our personography as a war hero appears inside an advertisement within a magazine published after the date of the war in question, the advertiser may be using that person's identity to promote the goods being advertised"; "people whose names appear in both advertisements and poetic dedications command greater social capital than those whose names appear in either genre alone"). And the conceptual leap from complex statements of this kind to larger conclusions is greater still.

This paper seeks to explore how the modeling of textual data for humanities research connects to the high-level research questions humanities scholars address in their scholarly writing. It offers a detailed description of the modeling (transcription, text encoding, metadata) and the high-level research goals for two closely related digital collections: the Early Caribbean Digital Archive and the Women Writers Project. It then traces critically the inferential steps by which we seek to get from the data being captured to the theoretical concepts ("culture", "geography", "influence", etc.) that animate the research. The goal of the paper is to provide a much more exacting and thorough understanding of the complexity of data modeling required to support the argumentative nuance and conceptual subtlety of real-world, high-quality humanities research.

In particular, our focus is on the possibilities and challenges of knowledge production afforded by modeling and encoding archives of materials that concern marginalized persons and non-canonical texts and histories. The two projects under discussion here are engaged in bringing to visibility texts and narratives that had previously been submerged beneath (or concealed within) more culturally prominent discursive forms. The Women Writers Project is currently engaged in a grant-funded collaborative research project funded by the National Endowment for the Humanities titled "Cultures of Reception" (<http://www.wwp.brown.edu/research/projects/reception/>) which gathers and digitizes periodical reviews of late 18th- and early 19th-century women's writing, to support the study of patterns of reception in an emerging transatlantic literary culture. The Early Caribbean Digital Archive (<http://www.northeastern.edu/nulab/the-early-caribbean-digital-archive/>) is digitizing a variety of Caribbean textual sources from the same period, with special emphasis on the study of the emerging culture of commodity circulation and its relation to the transatlantic slave trade and submerged narratives of race and gender. Both projects involve detailed TEI encoding of textual sources animated by research goals such as these:

- trace and map the relations between texts as a function of time, human agency, and geography
- bring into visibility relations between locations of print activity across the Caribbean archipelago
- show relations among individuals, such as printers, consumers, merchants, runaway slaves, missionaries, plantation owners, abolitionists, military figures, and colonial political figures
- map relations between legislation, commodity prices, geography
- map the geographic circulation of literary tropes
- trace changes in the culture of reviewing over time with respect to the emergence of a transatlantic literary culture
- bring to visibility the evaluative frames of reference within which women's writing is read
- trace the cultural frame of reference for reviewers in England and in North America

In both cases, the projects must make conceptual and inferential bridges between the specific assertions constituted in the markup (observations about genre, named entities, time and location, textual structure, references to circulating cultural objects such as commodities and texts) and concepts operating at a much more abstract level: "culture", "geography", "influence", "relations", "frames of reference." Humanities scholars are comfortable using terms like these in their writing, as the currency of methodologies from cultural studies to cultural

geography indicates, but what forms of evidence and inferential reasoning do they entail at the level of textual markup?

We can unpack here, in a preliminary way, the kinds of reasoning through which these bridges might be built, and the final paper will explore these in more detail. First, there is a set of direct modeling activities (transcription and markup) through which the texts are constituted as research evidence. The activities of transcription produce from the source document assertions about the existence of strings of characters, and markup allows the transcriber and editor to identify areas where this evidence is ambiguous or missing. Through markup the editors can also identify specific strings as references to certain types of named entities (persons, places, books, publishing houses, ships, shipping companies, legislative bodies, etc.) and can associate these references with their target to provide unambiguous entity identification via linked data authority records. At a structural level, the markup can also be used to identify the genre and format of texts and parts of texts, and to associate metadata (author, date and place of production, etc.) with these. Following these activities, we must venture to make inferences based on our modeling. For example, the markup allows us to give greater precision to inferences common in text analysis: instead of judging collocation based on raw word proximity, we can identify word pairs or groups as being within the same textual feature (paragraph, poem, letter, advertisement, heading). In specific cases, we may be able to infer something more from such collocation, such as a connection between two authors mentioned in the same paragraph of a review. Genre and format information may enable us to sharpen these inferences further: mentioning an author in a review means something different from mentioning an author in a dedication; a commodity carries a different cultural freight when listed in an advertisement, a bill of lading, a receipt, a letter, a legislative document. The name of an enslaved person means something different when it appears in a bill of sale, a runaway slave notice, a poem. Taking metadata into account, we can also localize documents in space and time, which gives us the possibility of (cautiously) identifying trends, causation, cultural significance.

In building these bridges, we need to be attentive to the gaps or weak inferential points as we move from the modeling to the research. For example, what does it mean to infer relationships between entities from their proximity within documents? Where are these inferences strongly grounded (for instance, co-authorship reflected in metadata records) and where are they weak (for instance, two names appearing in the same paragraph of a historical account)? Or, in another vein, is the model of geography that emerges from textual attestation (i.e. the inventory of place names and location references) adequate for the kinds of geographical analysis we want to do? Or again, what does "circulation" (whether of physical objects or of ideas) look like as attested in data of this kind and how do we discover it? Finally, what is the relation between the encoding categories we have identified here and our knowledge production with respect to marginalized texts, persons, and narratives? How will our modelling decisions erase or repeat historical occlusions in the archive, not only by determining what aspects of the text are marked but also by imposing existing frames of knowledge on the archive?

References

1. Bauman, Syd, David Hoover, Karina Van Dalen-Oskam, and Wendell Piez (2012). *Text Analysis Meets Text Encoding*. Panel session, DH2012, University of Hamburg, July 2012. www.dh2012.uni-hamburg.de/conference/programme/abstracts/text-analysis-meets-text-encoding
2. Flanders, Julia and Matthew L. Jockers. (2013). *A Matter of Scale*. Keynote Lecture from the Boston Area Days of Digital Humanities Conference. Northeastern University, Boston, MA. March 18, 2013. digitalcommons.unl.edu/englishfacpubs/106.
3. Our Cultural Commonwealth (2006). *The report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences*. ACLS.

4. Huitfeldt, Claus, C. M. Sperberg-McQueen, and Allen Renear (2001). *Meaning and Interpretation of Markup*. Markup Languages: Theory & Practice 2.3: 215-234.

Monolith: Materialised Bits, the Digital Rosetta Film

Fornaro, Peter

peter.fornaro@unibas.ch
Digital Humanities Lab

Wassmer, Andreas

a.wassmer@unibas.ch
Digital Humanities Lab

Rosenthaler, Lukas

lukas.rosenthaler@unibas.ch
Digital Humanities Lab

Gschwind, Rudolf

rudolf.gschwind@unibas.ch
Digital Humanities Lab

Introduction

Digital storage systems are like tins. They might have some content but the only evidence are labels, captions or any kind of lettering. In addition maybe the weight of the tin could be taken to judge about its inside but still to get full assurance the tin has to be opened to be able to identify the content correctly. In order to open it an appropriate tool is necessary and another one to take something out of it, something like a fork or a spoon.

A digital storage like a hard drive behaves very similar to that but in case of magnetic recording the data is not only invisible from in and outside the storage, not even can it be detected physically by a human being. Digital data has no appearance that can be touched, we have no sense for it.

In addition to that any bit stream stored on a data carrier needs to be migrated after a certain time to ensure accessibility and consistency because following three factors endanger the digital archiving process:

- The storage media decays over time and it can fail by aging.
- Hardware gets incompatible so that accessing the data becomes impossible.
- File formats – technical metadata – change and develop over time. Therefore programs to interpret the file content might not be available in future.
- Missing contextual metadata make any digital bit stream more or less useless.

Migration

Each of these possibilities is a major drawback for cultural heritage preservation and each one renders data into digital waste, data that is either lost completely or without meaningful sense. Continuous migration and copy storage content in periodic intervals are today's best practice to transform binary information into the future. Theoretically migration works very well because digital data can be copied without loss, which is one of the major advantages of any digital code like binary information or our even alphabet. The ability to copy data lossless allows not only the arbitrary replacement of the data carrier, it also allows to increase redundancy by storing multiple copies, as e.g. proposed by LOCKSS¹. The down side of migration is the financial effort associated with it. Independent of the specific costs per migration, archiving is getting expensive sooner or later because of the short lifetime of the technology. In addition its dependence on numerous cascaded technologies makes it a fragile process that can cause dramatic data loss if only one of the incorporated components fails.

Migration can be omitted if the storage media fulfills the following requirements:

- It must contain human readable metadata in order to describe the archived object and its context.
- Information on how to recover the original file (the decoding manual) must be part of the metadata. This knowledge is the key to interpret the archived byte stream.
- The file format must be well documented (open format) and it must be widely used to ensure its accessibility over time.
- Digital data is stored hardware independent as far as possible. Thus it is not affected by the change of technology.

If a medium claims to be suitable for long-term preservation of digital data it has to fulfill more requirements. Lunt et al.² identified 7 characteristics, which are particularly interesting to archivists regarding preservation of digital data. The first says, there shouldn't be active maintenance or migration required to preserve actual data. They continue with: 2) no special storage conditions are necessary to preserve the storage media; 3) a minimum lifetime of at least 100 years, preferably more should be supported; 4) no energy is required to maintain the data; 5) the media is easily transported; 6) the data format is widely adopted; 7) the medium has a large storage capacity.

Bits-on-Film Approach

Facing those facts the Digital Humanities Lab of the University of Basel has developed a workflow for migration-less preservation of digital data on optical media called "Monolith". It combines the advantages of photographic material and standard digital imaging technology to create a long-term migration-less archiving system. This is achieved by the hybrid characteristics of the optical carrier. Any arbitrary binary bit stream is put right besides human readable technical, structural, and contextual metadata. Original files of any appropriate format are stored on film as visual 2D-barcodes. Technically spoken every bit of the original bit stream (the file to be stored) is converted into a spot representation on film. A full bit stream then results in a two-dimensional image, an "image of bits". In other words the logical data-bits are transformed and represented by dye or silver of photographic film. This process can be regarded as a materialization of binary data, which becomes visual and physical. Monolith has no limitation regarding the format of the file to be archived. However, the documentation of the file format must be part of the metadata and therefor it should be an open standard like the widely used PDF-A or image formats like JPEG2000⁽³⁾. Metadata can be stored binary or as human readable text information, e.g. encoded and written in letters on film as any of the well-known standards like Dublin Core, METS or others.

This approach has various advantages: First, and most important, the bit stream on film can be read/captured by any digital camera, there is no special hardware necessary to transform the physical representation of the bits back into logical states within the computer system. This can be compared to the process of seeing. As human beings see letters – in fact digital data – the camera sees signs, spots on film – binary data; Monolith has a visual interface. The decoding of the binary bit stream is well defined because the explanation of the code is an inseparable part of the technical metadata set written on film. Like no other storage media Monolith can not only contain barcodes and text but images – e.g. thumbnails – as well. Besides its technological features the storage film has another advantage. It can be stored the same way as regular archival film. There are no special storage conditions necessary nor does the film need any specific care. Therefore, Monolith can be regarded more as an "engraved stone" than as a data storage for computer systems. It is a "Digital Rosetta Film".

But is the application of optical film for archival purposes reasonable these days? Many companies stopped production high fidelity film material and very likely the quality – not stability – of film will drop in the future. For the representation of photographic images this is of course a major draw back since image quality is directly related to film quality. In case of Monolith this is irrelevant. The only function the film has to fulfill is to separate dots spatially, requirements that are achieved

by most photographic materials. The quality of digital originals will not be disturbed by film quality because they are stored as binary data and therefore decay of the material has little impact. In addition any well-known error correction method can be applied. The concept of a binary representation of data is a simple but a very efficient solution and it is the reason why every computer storage system is adapting this concept⁴. Even if film won't be available in future, for any existing Monolith this means no impact. Not for the sustainability nor for the future ability for data recovery.

All those features show that materialized bits are not only a nice concept to mimic historic documents but also an efficient way to transport digital cultural heritage into the future⁵. In the presentation we will show how Monolith works and what its advantages are.

Monolith™ on 35mm color material

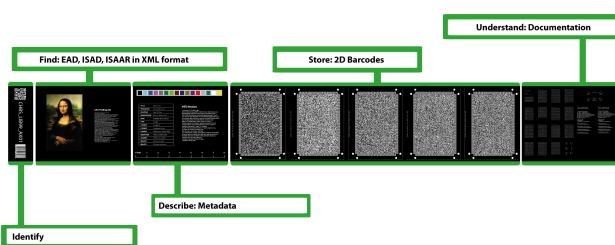


Fig. 1: Monolith™ includes all necessary information for future information recovery. Especially contextual metadata and the decoding manual to understand the structure of the bit-pattern.

Conclusion

Monolith is a solution that has made its way from university to a commercial company. It shows that there is a possibility for an alternative solution for classical digital archiving, that doesn't need to be migrated. The advantages are not only of technical but also of economical nature. Even if costs for plain storage media decrease with time, total costs of ownership for archived digital data increased in the last years continuously and migration is the primary costs driver of archiving. Therefore Monolith can not only be an answer for technological but also for economical challenges on the way of digital information to the future.

References

1. Vicky Reich & David S.H. Rosenthal. *LOCKSS (Lots Of Copies Keep Stuff Safe)*, Presented at Preservation 2000: An International Conference on the Preservation and Long Term Accessibility of Digital Materials, December 7-8, 2000, York, England. Also published in *The New Review of Academic Librarianship*, vol. 6, no. 1, 2000, pp. 155-161. doi:10.1080/13614530009516806
2. Barry M. Lunt, Matthew R. Linford, Robert Davies (2012), *Research on Another Permanent Data Storage Solution*, Proc. Archiving 2012, IS&T, pg. 19-21
- A list of file formats suited for archiving can be found at www.kost-ceco.ch/wiki/whelp/KaD/index.php.
4. Shaw Rodney, Selected Reading in Image Evaluation, SPSE, ISBN 0-89208-085-X
5. Florian Müller, Peter Fornaro, Lukas Rosenthaler, Rudolf Gschwind (2010) *PEVIAR: Digital Originals*, ACM Journal on Computing and Cultural Heritage, Volume 3, Issue 1. ACM 2010

Beyond Style: Literary Capitalism and the Publishing Industry

Fuller, Simon

simonfuller9@gmail.com

National University of Ireland, Maynooth

O'Sullivan, James

josullivan.c@gmail.com

Pennsylvania State University / University College Cork

1. Reinvention of the Publishing Industry

Having developed his commercial expertise as an advertising executive, James Patterson went on to forge a publishing empire that, according Forbes, earns him in the region of \$91 million each year.¹ Leaving his position as an adman to become a full-time writer in 1996,² a decade later his work had already grossed \$1 billion.³ To date, his name has appeared on over 100 books that have sold a combined 300 million copies.⁴ Patterson is the driving force behind the approach to marketing that has allowed him to overtake his competitors at the top of the bestseller lists, from dictating the advertising campaigns for his titles, to the manner by which they are distributed.⁵ Patterson's method is described as "a literary assembly line",⁶ an observation which he seemingly embraces: "I look at it the way Henry Ford would look at it," is his response to criticisms of his approach.⁷ His vision is such that Little, Brown, Patterson's publisher, has restructured its organisation in an effort to meet the author's requirements. In a lengthy piece for *The New York Times*, Jonathan Mahler quotes former Little, Brown publisher, Sarah Crichton: "To have one writer really start needing, and even demanding, the lion's share of energy and attention was difficult. There were times when some of us resented that. When Jim felt that resentment, he roared back. And he was too powerful to ignore."⁸ In his case study of Patterson's marketing methods, John Deighton outlines how the author commissioned his own studies in an effort to identify what potential readers want from a novel, so that he could deliver his titles to the widest possible audience: "The Little, Brown publishing group recognized that Patterson was a most unusual author, one who could teach them a lot about selling."⁹ Patterson is, according to Time-Warner Publishing's Larry Kirshbaum, "the first real brand-managed author".¹⁰

As a reader himself, critically acclaimed writers such as James Joyce and Gabriel García Márquez are amongst Patterson's preferred authors.¹¹ Patterson admits that there is a distinction between the aforementioned literary works and his novels: "These books are entertainments. It's a very different process than if you're trying to write *Moby-Dick*, or *The Corrections*. That's painful. That's different from very simple, plot-oriented storytelling."¹² Patterson is frequently criticised for his approach, but writing in the fashion of Joyce or Marquez, is a task which he suggests is beyond his ability as an author: "After reading *Ulysses*, I knew I couldn't write anything that great. I don't have it in me."¹³ Thus, producing a high volume of entertaining novels has been his objective.

2. Patterson's Collaborative Process

To achieve the prolific output that we see today, Patterson enlists the support of numerous collaborators: "It was Patterson who first showed that television advertising could work for books. More radically, he has demonstrated that working with co-writers can dramatically multiply sales."¹⁴ Patterson sees little difference between his approach to collaborative writing and those practices that have long been central to other sectors within the culture industry: "It isn't terribly groundbreaking ... The newspaper business, the movie business—they're full of teams. A lot of art was done by teams ..."¹⁵ "My short answer to the question as to why work with other people is Gilbert and Sullivan, Rodgers and Hammerstein, Woodward and Bernstein, Lennon and McCartney and it goes on," he offered in *The Guardian*.

Patterson is quite forthright when detailing his collaborative process: "I'll write an elaborate outline, maybe 70 pages, very detailed, clear, and focused. The co-author will write the first draft, and I'll see the work every few weeks. I'll do two to seven

more drafts.¹⁶ Described as "a natural born writer", many of Patterson's collaborators have used the opportunity as a platform from which to launch their solo careers.¹⁷ Gaby Wood comments in *The Guardian* that the sentences in Patterson's novel "are not designed to be lingered over", "[t]hey are more or less all plot".¹⁸ But, while it seems that Patterson revises each chapter as they are drafted, as well as making all of the final revisions alone,¹⁹ the extent to which he contributes authorial material directly, in terms of writing, remains unclear. Patterson has stated that his "name on the cover is the assurance of a good read"²⁰ – using computational methods central to Digital Humanities scholarship, this paper seeks to determine the volume of writing that Patterson contributes before offering such an assurance.

3. Methodology & Results

We evaluated the relative contributions of Patterson and his collaborators using versions of Burrow's Delta, a widely-used lexical measure for English texts.²¹ We selected the collaborators Peter de Jonge and Andrew Gross for this investigation. Patterson, by his own account, allocates most of the actual writing to his junior partners. It was in such a respect that we formed the working hypothesis that the collaborative works would be stylistically more similar to texts written primarily by Patterson's co-authors, than to any of the novels attributed to Patterson alone. We expected the lexical features that we employed to pick out the primary writer. This would correlate with much of the field's existing research, for instance by Patrick Juola,²² who suggests that, in attempted forgeries, the lexical signature of the forger overrides the semantic content which might associate it with the impersonated party. However, this is not a foregone conclusion. Jan Rybicki has demonstrated that in literary translations, the original authorial signals dominate that of the translator when using Burrow's Delta cluster analysis. In other words, the semantic imprint survives the translation process although all of the lexical features examined are written by the translator.²³ Our second hypothesis was that, given the former, Patterson's contribution would be strongest at critical moments in the text. Given the plot-driven genre, we believed that these would typically be present at the beginning and end of the novels.

To test our first hypothesis, we employ a "bootstrap consensus tree" cluster analysis over maximum frequency words ranging from 100 to 1000, in intervals of 100, with the Burrow's Delta metric, using the "stylo" package for R.^{24 25} We avail of a consensus strength of 0.5, meaning that we formed a tree showing proximity wherever this occurred in 50% or more of the 10 maximum frequency clusterings described.^{26 27} For our second hypothesis, we use the Rolling Delta technique.²⁸

^{29 30 31} To provide a general intuitive description of this method, Burrow's Delta distances are measured between the collaborative text and single-author texts for each participating author. However, distances are measured to "windows" of the collaborative text, allowing for estimation as to which sections carry the stylistic fingerprint of one contributor over another. Sample single-author tests are then plotted over the baseline of the collaborative text, where greater proximity to the baseline indicates greater stylistic similarity, as defined by the delta distance metric.

In this paper, we examine the following collaborative texts:

Patterson & De Jonge:

- Beach House (2003)
- Beach Road (2006)

Patterson & Gross:

- 2nd Chance (2002)
- 3rd Degree (2004)
- Judge and Jury (2006)

Our solo texts, by author, are as follows:

DeJonge:

- Shadows Still Remain (2009)
- Buried On Avenue B (2012)

Gross:

- The Dark Tide (2008)
- Don't Look Twice (2009)
- Killing Hour (2011)
- 15 Seconds (2012)
- No Way Back (2013)

For Patterson, we used this fixed set of nine solo works:

- First to Die (2001)
- Four Blind Mice (2002)
- The Lake House (2003)
- London Bridges (2004)
- Maximum Ride: The Angel Experiment (2005)
- Maximum Ride: Saving The World And Other Extreme Sports (2007)
- I, Alex Cross (2009)
- Fang (2010)
- Nevermore (2012)

The following visualisation displays our bootstrap consensus tree over the entire dataset:

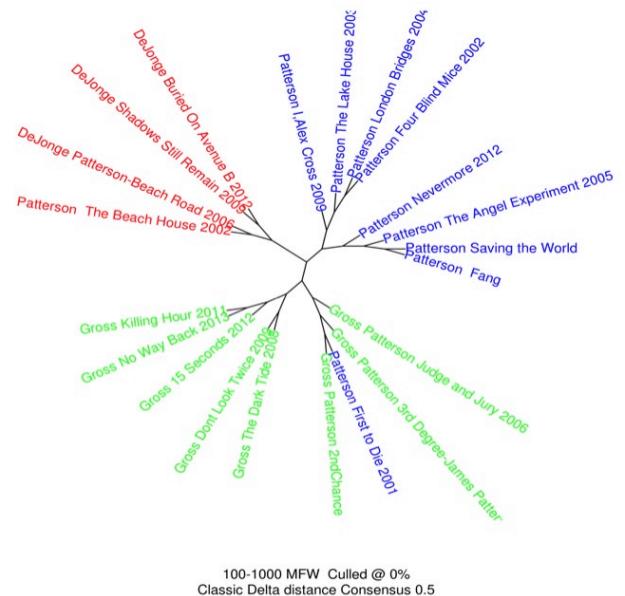


Fig. 1: Bootstrap consensus tree

As predicted, the collaborative works all cluster with the respective junior writer. Within both the De Jonge and Gross clusters, the collaborative works form a distinct sub-cluster. Within the Patterson cluster, the *Maximum Ride* series of novels are separated from another cluster consisting of *Alex Cross* novels and the Patterson novel, *The Lake House*. One surprise result is that *First to Die*, a solo Patterson text, is clustered with the subsequent works in the *Women's Murder Club* series, which he wrote with Andrew Gross. This could simply represent a limitation of the delta metric over these texts, or alternatively, it could indicate that Gross was so influenced by the particular style that Patterson manifested in this work that he imitated it more exactly than Patterson managed in any of the other works under examination. We discount a third possibility, that the new collaborative series, *Women's Murder Club*, opened with a solo Patterson work to kick-start sales. While such an interpretation would align with the marketing ingenuity of both Patterson and Gross, it is without sufficient empirical foundation.

Our full study comprised rolling deltas for all collaborative texts, under a number of different setting. For the purpose of this abstract we include just two rolling delta studies, *First to Die* and its sequel in the series, *Second Chance*, respectively, both with pronouns deleted from an initial most frequent word count of 1,000:

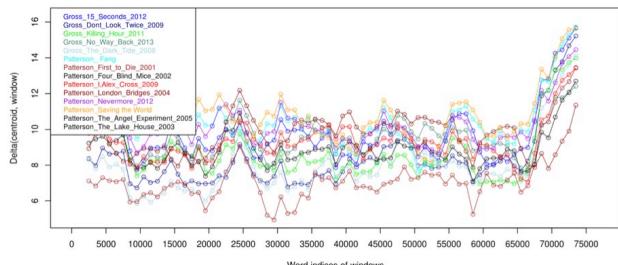


Fig. 2: Rolling Delta analysis

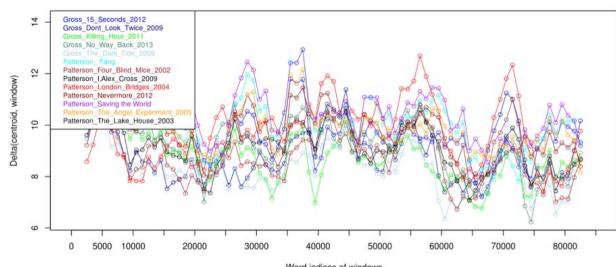


Fig. 3: Rolling Delta analysis

For *Second Chance*, Gross' texts are closest throughout, apart from *First to Die*, which, as we have already discussed, is attributed solely to Patterson. The model in *First to Die* is more interesting as the work appears like a true collaboration in which the authors have shared the task of writing passages or sentences with Patterson intervening at critical junctures.

4. Conclusions

The quantitative data suggests that Patterson's collaborators perform the vast majority of the actual writing. Therefore it seems that, unlike translation, the semantic signal from Patterson is dominated by the lexical signal of the other writer. We think the explanation for this difference is the density of the semantic message in each case – when implementing the outline of a general plot, there is more freedom for sentence and passage construction.

A full stylometric study of Patterson would also need to measure his contribution to the abstract entity called the plot in the works under examination. As Patterson says: “above all my brand stands for story. I became successful when I stopped writing sentences and started writing stories. Editors think it's about style. It's not. It's all story”.³² On the one hand, we note that for Patterson, this is just as well, since our analysis shows that his stylometric fingerprint is sometimes weak, even in his solo works. On the other hand, we recall Aristotle also believed, in his analysis of tragedy, that plot ($\mu\bar{\theta}\delta\sigma$) - the “arrangement of the incidents ($\eta\tau\omegaν\piραγμάτων\sigmaύστασις$)” is the most important element of the work.³³

In the *Arcades Project*, Walter Benjamin collects several references to writers who brought industrial methods of production to bear upon the process of literary creation in 19th Century France. While the 19th Century “witnessed an institutionalization of the split between technology and art to a degree previously unknown in history”.³⁴ Benjamin pays particular attention to those who found new ways of conjoining the two, such as Eugene Scribe, popularizer of the “well made play (piece bien faite)”, and Alexandre Dumas, who both industrialised the writing process. Dumas was contemporaneously described as running a “factory of novels,”³⁵ and, like Patterson, simultaneously worked on multiple novels, with an output of 400 novels and 35 dramas in 20 years.³⁶ Lucas-Dubreton recounts how accusations of plagiarism by one De Mirecourt’s eventually led to Dumas publicly recognising his co-authors, and arranging better terms of payment for them.³⁷

In language which foreshadows the descriptions of Patterson as a “brand-managed” author, Lucas-Dubreton caricatures the

allegations against Dumas thus: “Dumas did not exist at all, he was only a myth, a trademark invented by a syndicate of editors to dupe the public”.³⁸

Patterson and Dumas employ modes of authorial production which echo the economic advancements of their times. Unlike Dumas, Patterson has never been questioned in relation to his collaborative process, and such anomalies as we have detected probably indicate most of all the immaturity of stylometry as a method. Regardless, it seems Patterson, an epitome of benign capitalism, has offered sufficient accreditation, tutelage and financial reward for those who work for him.

In this paper, we will discuss our analysis in the context of these publishing practices, both new and old, and how literary collaborations might be approached from a stylometric perspective, using Patterson and his reinvention of the industry as a useful test case.

References

1. **Bercovici, Jeff.** *The World's Top-Earning Authors: With '50 Shades,' E.L. James Debuts At No. 1.* Forbes. Web. 6 Oct. 2013.
2. **Mahler, Jonathan.** (2010) *James Patterson Inc.* The New York Times 24 Jan 2010. NYTimes.com. Web. 14 Aug. 2013.
3. **Wroe, Nicholas.** (2013) *James Patterson: a Life in Writing.* The Guardian. 11 May 2013. Web. 19 Aug. 2013
4. **Wroe, Nicholas.** (2013) *James Patterson: a Life in Writing.* The Guardian. 11 May 2013. Web. 19 Aug. 2013
5. **Mahler, Jonathan.** (2013) *James Patterson Inc.* The New York Times 24 Jan. 2010. NYTimes.com. Web. 14 Aug. 2013
6. **Deighton, John.** (2006) *Marketing James Patterson.* Case Study. Boston. Harvard Business Publishing. Web. 24 August 2013. pp 4.
7. **Deighton, John.** (2006) *Marketing James Patterson.* Case Study. Boston. Harvard Business Publishing. Web. 24 August 2013. pp 1.
8. **Mahler, Jonathan.** (2010) *James Patterson Inc.* The New York Times 24 Jan. 2010. NYTimes.com. Web. 14 Aug. 2013
9. **Deighton, John.** (2006) *Marketing James Patterson.* Case Study. Boston. Harvard Business Publishing. Web. 24 August 2013. pp 5.
10. **Deighton, John.** (2006) *Marketing James Patterson.* Case Study. Boston. Harvard Business Publishing. Web. 24 August 2013. pp 5.
11. **Mahler, Jonathan.** (2010) *James Patterson Inc.* The New York Times 24 Jan 2010. NYTimes.com. Web. 14 Aug. 2013.
12. **Flood, Alison.** (2010) *James Patterson Brings in \$70m to Become World's Highest-earning Author.* The Guardian, 20 august 2010. Web. 17 Aug. 2013.
13. **Patterson, James.** (2012) *Life's Work: James Patterson.* Harvard Business Review. Web.
14. **Wroe, Nicholas.** (2013) *James Patterson: a Life in Writing.* The Guardian. 11 May 2013. Web. 19 Aug.
15. **Patterson, James.** (2012) *Life's Work: James Patterson.* Harvard Business Review. Web.
16. **Patterson, James.** (2012) *Life's Work: James Patterson.* Harvard Business Review. Web.
17. **Belena, Ruth.** (2013) *How James Patterson Has Launched the Careers of His Co-authors.* Helium. 5 June 2013. Web. 16 Aug. 2013.
18. **Wood, Gaby.** (2009) *The World's No 1 Bestseller.* The Guardian. 5 Apr. 2009. Web. 17 Aug. 2013.
19. **Belena, Ruth.** (2013) *How James Patterson Has Launched the Careers of His Co-authors.* Helium. 5 June 2013. Web. 16 Aug. 2013.
20. **Deighton, John.** *Marketing James Patterson.* Case Study. Boston. Harvard Business Publishing, 2006. Web. 24 August 2013. pp 2.
21. **Burrows, J. F.** (2002) *Delta: A Measure of Stylistic Difference and a Guide to Likely Authorship.* Literary and Linguistic Computing 17: 267-287
22. **see e.g. Juola, Patrick.** (2013) *Stylometric Report Heartland Institute Memo*, retrieved from wattsupwiththat.files.wordpress.com/2012/03/memoreport.pdf 18th October 2013.

23. **Rybicki J** (2012) *The great mystery of the (almost) invisible translator: stylometry in translation*. In: Oakes M, Ji M, editors. Quantitative Methods in Corpus-Based Translation Studies. Amsterdam: John Benjamin.

24. **Eder, M., and J. Rybicki.** (2011) *Do Birds of a Feather Really Flock Together, or How to Choose Test Samples for Authorship Attribution*. Stanford: Digital Humanities.

25. **Eder, Maciej, Mike Kestemont and Jan Rybicki.** (2013) *Stylometry with R: a suite of tools*. Digital Humanities 2013: Conference Abstracts. University of Nebraska-Lincoln, Lincoln. pp. 487-89.

26. **Wilkinson** (1996), *M. Majority-rule reduced consensus trees and their use in bootstrapping*. Molecular Biology and Evolution 13(3):437-44.

27. **Paradis, E. and Claude, J. and Strimmer K** (2004) *APE: analyses of phylogenetics and evolution in R* Bioinformatics 20 (2): 289-290.

28. **Rybicki, Jan, Mike Kestemont and David Hoover.** (2013) *Collaborative authorship: Conrad, Ford and rolling Delta*. Digital Humanities 2013: Conference Abstracts. University of Nebraska-Lincoln, Lincoln. pp 368-71.

29. **Rybicki, J., Kestemont, M. and Hoover D.** (2013). *Collaborative authorship: Conrad, Ford and rolling delta*. In: "Digital Humanities 2013: Conference Abstracts. University of Nebraska-Lincoln, Lincoln, NE, pp. 368-71.

30. **Hoover, D.** (2011). *The Tutor's Story: a case study of mixed authorship*. In: "Digital Humanities 2011: Conference Abstracts". Stanford University, Stanford, CA, pp. 149-51.

31. **Dalen-Oskam, K. van and Zundert, J. van** (2007). *Delta for Middle Dutch: author and copyist distinction* in Walewin. Literary and Linguistic Computing", 22(3): 345-62.

32. **Deighton, John.** (2006) *Marketing James Patterson*. Case Study. Boston. Harvard Business Publishing. Web. 24 August 2013. pp 5.

33. **Aristotle** (1932). *Poetics* 1450a. Trans. W.H. Fyfe. Cambridge, MA, Harvard University Press; London, William Heinemann Ltd. From www.perseus.tufts.edu. Web 16 October 2013.

34. **Buck-Morss, Susan.** (1993) *The Dialectics of Seeing*. Baskerville VA: MIT Press. pp. 126.

35. **de Mirecourt, Eugène. Alexandre Dumas (1845) & Co. Factory of Novels.** Cited in Benjamin, Walter. The Arcades Project. Trans. Eiland and McLaughlin. (2003) Harvard University Press [d3a,8] p. 749.

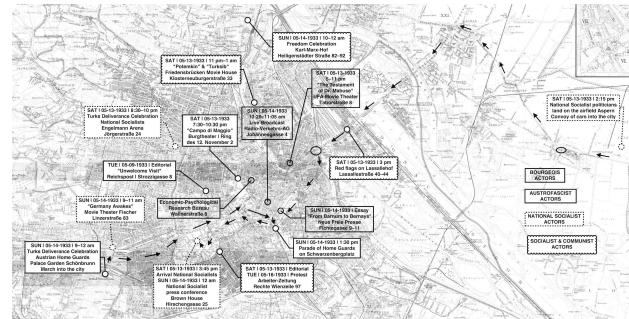
36. **Benjamin, Walter.** *The Arcades Project*. Trans. Eiland and McLaughlin. Harvard University Press (2003): [d4,2] p. 751-752. See also **Buck-Morss, Susan.** *The Dialectics of Seeing*. Baskerville VA: MIT Press (1993) especially pp. 136-142.

37. **Lucas-Dubreton, Jean.** The Fourth Musketeer: The Life of Alexander Dumas. Trans. Maida Castelhun Darnton. N. p. Web. 24 Oct. 2013

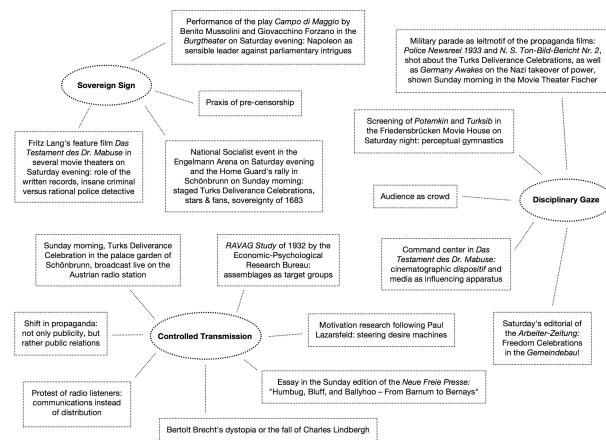
38. **Lucas-Dubreton, Jean.** *The Fourth Musketeer: The Life of Alexander Dumas*. Trans. Maida Castelhun Darnton. N. p. Web. 24 Oct. 2013. See also Benjamin, Walter. The Arcades Project. Trans. Eiland and McLaughlin. (2003) Harvard University Press: [P3a,3] p. 523.

Campus Medius investigates mediality as historical experience, focusing on a chronotope (Bachtin 2008 & Schlögel 2008) of twenty-four hours in Vienna between May 13 and 14, 1933, and presented on two web-based platforms: an interactive map and an interactive documentary. Methodologically, the project relates Bruno Latour's actor-network theory to dispositif analysis following Michel Foucault, directed at identifying the conditions under which the experiential field of modern media was able to emerge. This approach identifies three levels, distinct in terms of their perspective but empirically overlapping: an archaeology of knowledge forms, a genealogy of power relations, and a typology of subjectivation modes. The twenty-four hours under investigation are marked by the Turks Deliverance Celebrations (Ackerl 1984) held by the Austrian National Socialists and the Home Guards. These events, commemorating the 250th anniversary of Vienna's liberation from the Turkish siege in 1683, were oriented from the outset upon processes of mass communication: the rallies were prepared by the party-political press, partially broadcast live on radio, and captured in propaganda films. To create a counter-public sphere, the Social Democrats published programmatic editorials and organized open-air concerts in Vienna's municipal buildings. While the Burgtheater staged Benito Mussolini's play *Campo di Maggio*, the large cinemas were screening Fritz Lang's sound movie *Das Testament des Dr. Mabuse*, a film banned in Germany. Campus Medius reconstructs the media events of this time-space and traces them back to historical media *dispositifs*. Hence, the project consists of two integral parts:

1. Topography: an interactive map of the chronotope, including hypermedia documents such as video clips, newspaper articles, radio broadcasts, photographs, and archive files. On this mock-up, the events are arranged politically—socialist and communist actors, Austrofascist actors, National Socialist actors, and bourgeois actors.



2. Topology: an argumentative analysis of knowledge forms, power relations, and subjectivation modes underlying the exemplary time-space. Based on the concrete events, this interactive documentary examines functions of media in the classical, modern, and postmodern age—a periodization which constitutes the organizational structure as represented in the diagram.



CAMPUS MEDIUS--Topography and Topology of a Media Experience

Ganahl, Simon

Department for German Studies, University of Vienna, Austria

Solomon, Rory

Parsons The New School for Design, New York, USA

Brennan, Mallory

School of Media Studies at The New School in New York

Daftary, Darius

Lead Engineer at Artivest in New York

The technological architecture of the project is comprised of two platforms: the topography exists within the Urban Research Tool (URT) for geospatial mapping, developed at Parsons The New School for Design, and the topology exists within Zeega, a tool for interactive storytelling developed at Harvard's Sensory Ethnography Lab. These platforms provide two complementary lenses for exploring the collection of media material, and will be accessible through one interface: a website with the URL campusmedius.net. On its front page, an image of Vienna's historical map will be integrated with URT as a kind of basemap, rectified to align with the underlying OpenStreetMap data. The user can either wander through the actor-networks by clicking on the hypertextual points and paths, or follow the Zeega tours that offer three narrative montages, thereby allowing cross-references between the platforms. While the mapped topography describes the events and displays the related media in a networked structure, the Zeegas—titled Sovereign Sign, Disciplinary Gaze, and Controlled Transmission—attempt to explain the historical a priori of the actual experiences, corresponding to the periodization mentioned above. Apart from access to these tours, the interactive map as interface will provide links to three subordinate pages: a text that introduces the project to the users; a glossary of the main theoretical, topical, and technological terms; and a list of references and sources.

Age	Knowledge	Power	Subjectivation
classical	representation	sovereignty	God
modern	man	disciplinarity	individual
postmodern	communication	control	assemblage

The Turks Deliverance Celebrations connect the chronotope with the siege of Vienna in 1683 or—in a broader view—with the classical age. In Foucault's philosophy, this historical dispositif features representation as knowledge form, sovereignty as power relation, and God as subjectivation mode. (Foucault 1966, 1975 & 1976) As shown in the overview of the project's topology, the Viennese weekend actualizes not only these classical attributes, but also features of the modern dispositif in Foucault distinguished by man as knowledge form, disciplinarity as power relation, and individuals in masses as mode of subjectivation. (Foucault 1966, 1975 & 1976) Finally, the exemplary time-space also gives examples of a postmodern dispositif with communication as knowledge form, control as power relation, and assemblages as subjectivation mode. (Deleuze 1990, Hardt/Negri 2000 & Galloway/Thacker 2007) Campus Medius analyzes functions of media in these historical settings as well as effects of their actualization: sovereign leaders of the 17th and 18th century return as theatrical stars; movie houses resemble disciplinary institutions of the 19th century; controls of communication processes as established in the 20th century appear in the form of target groups and public relations; etc.

The research project deals with (a) representational, (b) methodological, and (c) empirical questions currently under discussion in cultural studies and humanities:

- a) Campus Medius is rooted in digital humanities. (McPherson 2009 & Gold 2012) In its clear division between topographical presentation (Presner 2010) and topological analysis (Deleuze 1986), the project utilizes two web-based platforms in order to harness each one's particular affordances. While the interactive URT map shows the actors through hypermedia documents, the Zeega tours focus on the emergence conditions of these events. The capacity of this narrative montages to provide an argumentative explanation of its subject is juxtaposed with the map's encouragement to explore the movie clips, newspaper articles, radio broadcasts, photographs, and archive files in various ways. This implementation sets a high value on technological and scholarly transparency: Campus Medius has been examined in a peer-review process, is being programmed with open-source software, and will be published open-access in the online journal *Sensate* in 2014.

b) Campus Medius contributes to the clarification of the relationship between actor-network theory (Latour 2005) and dispositif analysis (Foucault 1977). According to our thesis, the former suits the description of concrete events that involve human and non-human actors while the latter allows us to derive the examined cases from historical dispositifs of knowledge forms, power relations, and subjectivation modes. (Ganahl 2013) This association is implemented in the dual structure of topography and topology. Due to the media historical subject, the project not only takes up the philosophical (Deleuze 1989 & Agamben 2009) and sociological (Law 1992 & Bührmann/Schneider 2008) discussion, but also the so-called apparatus debate (Baudry 1975 & Rosen 1986).

c) Campus Medius relates actors that are usually covered in separate fields of study. Thus, the representative architecture of Schönbrunn Palace and the disciplinary character of Karl-Marx-Hof meet each other as scenes of antagonistic rallies. (Blau 1999) The sensible leader who Mussolini imagined as Napoleon in Campo di Maggio (Dietrich 1976 & Pyrah 2007) confronts the insane criminal as depicted in Lang's *Das Testament des Dr. Mabuse* (Jacques 1994 & Aurich 2001). While the chancellor's speech at the Turks Deliverance Celebration was broadcast live on radio, the bourgeois newspaper *Neue Freie Presse* ran an essay about Edward Bernays, a nephew of Sigmund Freud, who carried out propaganda as an "exact science" in order to direct public opinion via "group leaders." (Rundt 1933) Bernays had opened his New York office for public relations in the 1920s (Ewen 1996 & Tye 1998), but his concept to build lifestyles around products didn't fully thrive until Paul Lazarsfeld developed complementary statistical methods with his Economic-Psychological Research Bureau in Vienna. (Samuel 2010) The sociologist carried out a survey for Austrian radio in 1932 that broke with an understanding of the audience as a mass of individuals. (Mark 1996) By correlating the listeners' program wishes with their social data, he created target groups that could be exploited in commercial terms.

References

- Ackerl, I. (1984). *Die Türkenbefreiungsfeiern des Jahres 1933. Historische Jubiläen als politische Propagandavehikel*, Geschichte und Gegenwart, 1: 18–26.
- Agamben, G. (2009). *What is an apparatus? And Other Essays*. Stanford: Stanford University Press.
- Aurich, R. et al. (2001). *Fritz Lang. Leben und Werk. Bilder und Dokumente*. Berlin: jovis.
- Bachtin, M. M. (2008). *Chronotopos*. Trans. by Michael Dewey. Frankfurt a. M.: Suhrkamp.
- Baudry, J. L. (1975). *Le dispositif: approches métapsychologiques de l'impression de réalité*, Communications, 23: 56–72.
- Blau, E. (1999). *The Architecture of Red Vienna*. 1919–1934. Cambridge: MIT Press.
- Bührmann, A. D. & Schneider, W. (2008). *Vom Diskurs zum Dispositiv*. Eine Einführung in die Dispositivanalyse. Bielefeld: Transcript.
- Deleuze, G. (1986). *Foucault*. Paris: Minuit.
- Deleuze, G. (1990). *Post-scriptum sur les sociétés de contrôle*. In Deleuze, G. *Pourparlers*. Paris: Éd. de Minuit, pp. 240–47.
- Deleuze, G. (1989). *Qu'est-ce qu'un dispositif?* In Michel Foucault philosophie. Recontre internationale Paris 9, 10, 11 janvier 1988. Paris: Seuil, pp. 185–95.
- Dietrich, M. (1976). *Burgtheaterpublikum und Öffentlichkeit in der Ersten Republik*. In Dietrich, M. (ed.), *Das Burgtheater und sein Publikum*. Vienna: Verlag der Österreichischen Akademie der Wissenschaften, pp. 479–707.
- Ewen, S. (1996). *PR! A Social History of Spin*. New York: BasicBooks.
- Foucault, M. (1976). *La Volonté de savoir*. Paris: Gallimard (= Histoire de la sexualité, vol. 1).
- Foucault, M. (1977). *Le jeu de Michel Foucault*. In Foucault, M. (1994), *Dits et écrits*. Vol. III: 1976–1979. Ed. by Daniel Defert and François Ewald. Paris: Gallimard, pp. 298–329.
- Foucault, M. (1966). *Les Mots et les choses*. Paris: Gallimard.

- Foucault, M.** (1975). *Surveiller et punir. Naissance de la prison*. Paris: Gallimard.
- Galloway, A. R. & Thacker, E.** (2007). *The Exploit. A Theory of Networks*. Minneapolis: University of Minnesota Press.
- Ganahl, S.** (2013). *Ist Foucaults "dispositif" ein Akteur-Netzwerk?* In *foucaultblog*. www.fsw.uzh.ch/foucaultblog/blog/9/ist-foucaults-dispositif-ein-akteur-netzwerk
- Gold, M. K.** (2012). *Debates in the Digital Humanities*. Minneapolis: University of Minnesota Press.
- Hardt, M. & Negri, T.** (2000). *Empire*. Cambridge: Harvard University Press.
- Jacques, N.** (1994). *Das Testament des Dr. Mabuse*. Hamburg: Rogner & Bernhard.
- Latour, B.** (2005). *Reassembling the Social. An Introduction to Actor-Network-Theory*. New York: Oxford University Press.
- Law, J.** (1992). *Notes on the Theory of the Actor-Network: Ordering, Strategy and Heterogeneity, Systems Practice*, 5(4): pp. 379–93.
- Mark, D.** (1996). *Paul Lazarsfelds Wiener RAVAG-Studie 1932. Der Beginn der modernen Rundfunkforschung*. Vienna: Guthmann-Peterson.
- McPherson, T.** (2009). *Media Studies and the Digital Humanities*, Cinema Journal, 48(2): pp. 119–23.
- Presner, T.** (2010). *HyperCities: A Case Study for the Future of Scholarly Publishing*. In McGann, J. (ed.), *Online Humanities Scholarship: The Shape of Things to Come*. Texas: Rice University Press, pp. 143–54.
- Pyrah, R.** (2007). *The Burgtheater and Austrian Identity*. Theater and Cultural Politics in Vienna, 1918–38. London: Legenda.
- Rosen, P.** (1986). *Narrative, Apparatus, Ideology. A Film Theory Reader*. New York: Columbia University Press.
- Rundt, A.** (1933). *Humbug, Bluff und Ballyhoo. Von Barnum bis Bernays*, Neue Freie Presse (Vienna), May 14, 1933: pp. 25–6.
- Samuel, L. R.** (2010). *Freud on Madison Avenue. Motivation Research and Subliminal Advertising in America*. Philadelphia: University of Pennsylvania Press.
- Schlögel, K.** (2008). *Terror und Traum. Moskau 1937*. Munich: Hanser.
- Tye, L.** (1998). *The Father of Spin. Edward L. Bernays and the Birth of Public Relations*. New York: Crown.

The MAAYA Project: Multimedia Analysis and Access for Documentation and Decipherment of Maya Epigraphy

Gatica-Perez, Daniel

EPFL, Switzerland; gatica@idiap.ch
Idiap Research Institute, Switzerland

Pallan, Carlos

University of Bonn, Germany

Marchand-Maillet, Stephane

University of Geneva, Switzerland

Odobeza, Jean-Marc

EPFL, Switzerland;
Idiap Research Institute, Switzerland

Roman Rangel, Edgar

University of Geneva, Switzerland

Grube, Nikolai

University of Bonn, Germany

1. Introduction

Archaeology and epigraphy have made significant progress to decipher the hieroglyphic writings of the Ancient Maya,

which today can be found spread over space (in sites in Mexico and Central America and museums in the US and Europe) and media types (in stone, ceramics, and codices.) While the deciphering goal remains unfinished, technological advances in automatic analysis of digital images and large-scale information management systems are enabling the possibility to analyze, organize, and visualize hieroglyphic data that can ultimately support and accelerate the deciphering challenge.

We present an overview of the MAAYA project (<http://www.idiap.ch/project/maaya/>), an interdisciplinary effort integrating the work of epigraphists and computer scientists with three goals:

- (1) Design and development of computational tools for visual analysis and information management that effectively support the work of Maya hieroglyphic scholars;
- (2) Advancement of the state of Maya epigraphy through the coupling of expert knowledge and the use of these tools; and
- (3) Design and implementation of an online system that supports search and retrieval, annotation, and visualization tasks.

Our team approaches the above goals acknowledging that work needs to be conducted at multiple levels, including data preparation and modeling; epigraphic analysis; semi-automated and automated pattern analysis of visual and textual data; and information search, discovery, and visualization. In this abstract, we concisely describe three ongoing research threads, namely data sources and epigraphic analysis (Section 2), glyph visual analysis (Section 3), and data access and visualization (Section 4). We provide final remarks in Section 5.

2. Data sources and epigraphic analysis

The project focuses on Maya hieroglyphic inscriptions produced within the Yucatan Peninsula, inside the northern Maya lowlands, which encompasses sites within the Mexican states of Yucatan, Campeche, parts of Quintana Roo and a northern-most portion of Belize (see Fig. 1). Our research targets the three Maya Books (Codices) produced inside the Yucatan peninsula during the Postclassic period (1000–1521 AD). The first one is the Dresden Codex, housed at the University Library of Dresden, Germany. For this data source, our project relies on published facsimiles (Förstemann, 1880; Codex Dresden, 1962; Codex Dresden, 1989) and on high-resolution, open-access images provided by the SLUB. The Codex Madrid is stored at the *Museo de América* in Madrid, Spain, and for its study, our project relies on published facsimiles and line drawings (Codex Madrid, 1967; Villacorta and Villacorta, 1976). For the Paris Codex, the project relies on published facsimiles and images provided online by the *Bibliothèque Nationale de France*.

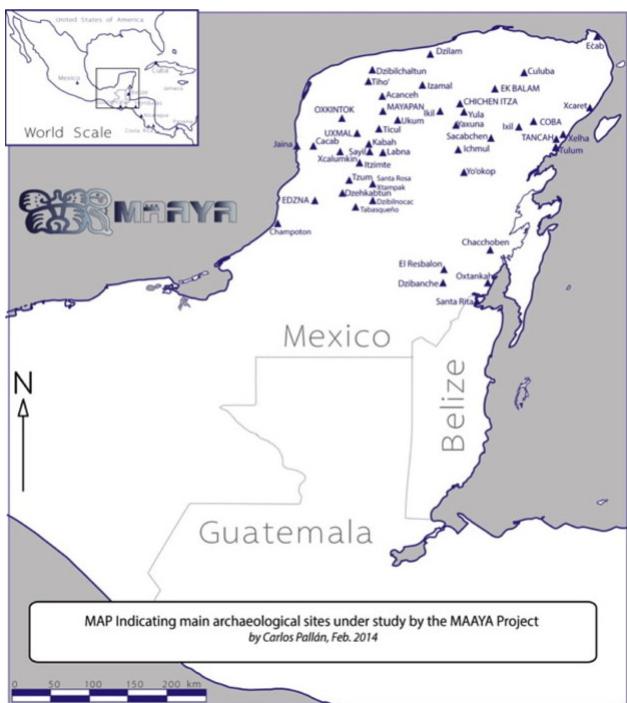


Fig. 1: Map indicating main archaeological sites under study by our project.

Codex pages were usually divided by red lines or *t’ols* (**Fig. 2**). Each of these *t’ols* is further subdivided in frames relevant to the specific dates, texts and imagery depicted. Frames contain several glyph blocks organized in a grid-like pattern with columns and rows, calendric glyphs, captions, and iconographic motives. Briefly stated, *t’ols* are “segmented” into their main constituent elements (**Fig. 2**). Images are post-processed and from these, high-quality, scale-independent vectorial images of the individual hieroglyphs and iconography are generated in three modes: (a) grayscale/color, (b) binary, and (c) reconstructed forms (marked in blue), which are based on epigraphic comparison of all available similar contexts (**Figs. 3-4**)

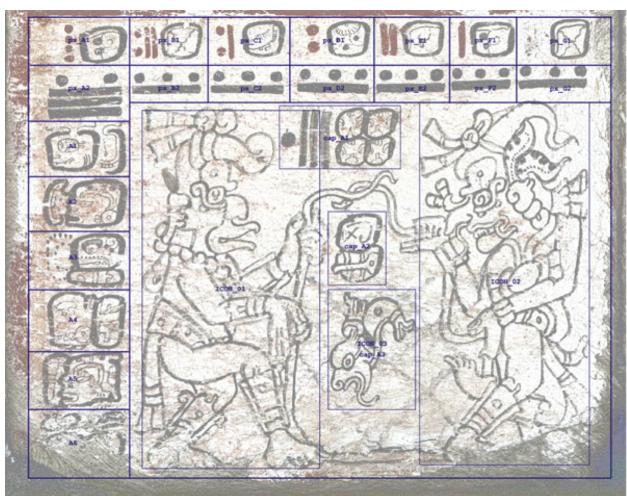


Fig. 2: Page 47c (44c) of the Dresden Codex framing main individual constituent elements (by Carlos Pallán based on SLUB online open source image)

The process of annotating the Codices entails an analysis comprising the following steps: (a) identification of individual signs on (Thompson, 1964) catalog, i.e. T0588:0181; (b) identification of individual signs on (Macri and Vail, 2008) catalog, i.e. SSL:ZU1; (c) identification of individual signs on (Evrenov et al., 1961) catalog, i.e. 400-010-030; (d)

identification of signs on (Zimmermann, 1956) catalog, i.e., Z0702-0060; (e) transcription, specifying phonetic values for individual signs as syllables (lowercase bold) or logograms (uppercase bold), i.e. **K'UH-OK-ki**; (f) transliteration, conveying reconstructed Classic Maya speech (words) formed by the combination of individual signs, i.e. k'uhul ook; (g) morphological segmentation, a division into morphemes for later linguistic analysis, i.e. k'uh-ul Ok; (h) morphological analysis, assigning each of the previous segments to a definite linguistic category, i.e. god-ADJ step(s)/foot; (i) English translation: "Divine step(s)/foot". Taken together, the processing steps within this workflow provide the ground for more advanced multimedia analyses (**Fig. 5**).

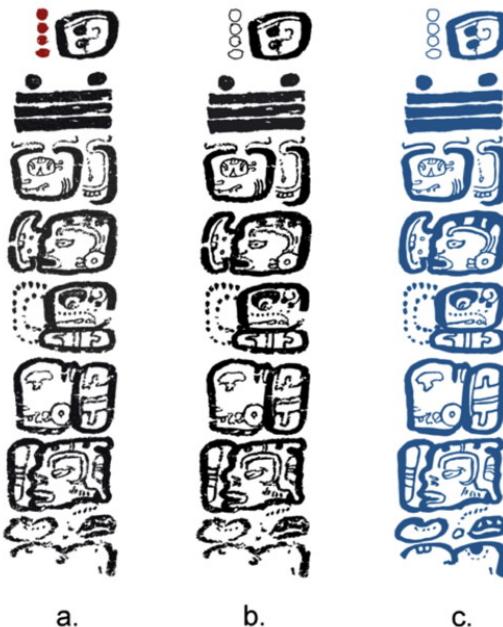


Fig. 3: Process to generate vectorial representations of the Dresden Codex: a) color/grayscale; b) binary; c) reconstructed (blue) forms (by Carlos Pallán based on SLUB online open source images)

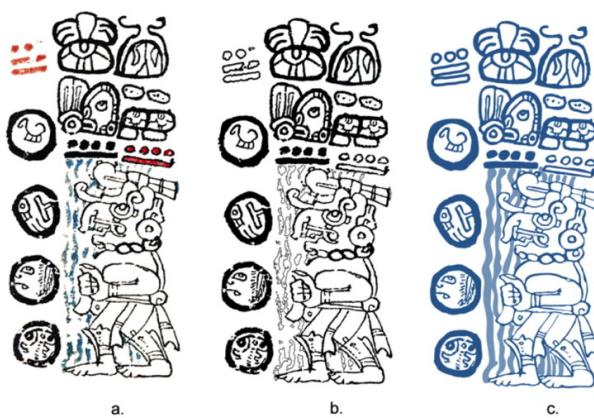


Fig. 4: Vectorial representations of the Madrid Codex, Page (T'ol) 10b, Frame 1: a) color/grayscale; b) binary; c) reconstructed (blue) forms (by Guido Krempel based on (Codex Madrid, 1967))

DRE	Page/T of 47c	Frame	1	1	1	1	1	1
Read. Order	O1	O2	O3	O4	O5	O6		
Collocation	A1	B1	A2	B2	A3	B3		
Glyph Date(s):								
Glyph-block								
Members (a, b, c...)	a, b	a, b	a, b, c	a, b	a, b	a, b, c		
Members Ranking a, b, c, ...	3, 3	3, 3	3, 3, 3	3, 3	3, 3	2, 1, 1		
Thompson	T0400-0010-0030	T0404-1057	T0033-0765b-0102	T0668-0102	T0044-1006b	T0162-0506-*0501		
Macri/Looper	SSL213	PC4	AMC-AP1182	M29-182	1M2-P68	32P-XH4-XKE2		
Evernor et al.	400-010-030	02-5-262	040-332-117	530-112	111-255	57A-S15-212		
Zimmermann 1956	20702-2060	Z-0130	20030-0707-0061	20169-0061	20080-0126	20139-1132-1121		
Transcription	WA'-ja	NAAH	K'UH-OK-ki	cha-ki	LEM?-NAL	TF-WAI*-HA*		
Transliteration	wo'j?/ wo'[h]aj?	nah[n]il	k'uh[ul] ok	Chaa[n]ik	Lem? Na!	t'i' waa* (t'i) ho*		
Segmentation	wo'j? / wo'[h]aj?	nah[n]il	k'uh[ul]	Chaa[n]ik	Lem? Na!	t'i' waa* / ho'		
Morph. analysis	stand-up-PAS	north	gpd-ADJ step(s) foot	rain.god	resplendent? maize god	mouthful tamale		
Translation	(he) raised/stood	(at) north	"the holy-step(s) Rain God"		[the resplendent?] maize god	"abundance of sustenance"		

Fig. 5: Multivariable fields used to annotate textual contents of Dresden Page (Tol) 47c (44c) (by Carlos Pallán)

3. Visual analysis of glyphs

Modeling Maya glyph shape is challenging due to the complexity and high intra-class variability of glyphs. We are developing methods to characterize glyphs for visual matching and retrieval tasks. In previous work, we proposed a shape descriptor based on visual bags-of-words (HOOSC: Histogram-of-Orientations-Shape Context) and used it for isolated glyph retrieval (Roman-Rangel, 2011). We are pursuing two research lines to extend our current capabilities.

Improved shape representations. Three directions are being considered: (1) the improvement of bag representations to retrieve syllabic glyphs. In particular, we developed a method to detect visual stop-words (Roman-Rangel, 2013a), and a statistical approach to construct robust bag-of-phrases (Roman-Rangel, 2013b); (2) the use of neural-network architectures like auto-encoders (Ngiam, 2011) that automatically build representations from training data. These approaches represent an alternative to handcrafted descriptors like the HOOSC, and provide a principled way to quickly adapt representations to different data sources (codex vs. monument glyphs); (3) the use of representations based on the decomposition of glyphs into graphs of segments, from which shape primitives can be extracted. This representation might be more suitable than histogram-based descriptors like HOOSC at identifying which strokes of a shape are discriminative, potentially allowing comparisons with so-called diagnostic features provided by epigraphers (Fig. 6).



Fig. 6: Three glyph instances of the same sign. Right: one diagnostic features and variant.

Co-occurrence modeling. We are exploring ways to exploit the fact that glyphs do not occur in isolation within inscriptions but in ordered groups (glyph-blocks) (Fig. 2). To this end, we are studying options to build models relying on glyph co-occurrence statistics or further accounting for the glyph spatial position within the blocks. We plan to investigate how such information can be used in a retrieval system to improve performance and to help scholars deal with unknown or damaged glyphs. This has several dimensions like query types (e.g. single glyphs with known identity of other glyphs within the block), and contextual combination of shape similarity with text metadata.

4. Data access and visualization

Our work in this direction focuses on visualization of and effective access to image databases with archaeological value. We are developing a repository that will serve further goals within the project. This database stores visual elements of the Madrid, Dresden, and Paris codices. It is complemented with an online system, shown in Fig. 7, which allows for capturing and annotation of codices. More specifically, the repository contains relevant information regarding the composition of the codices, such as hierarchical relations between components and bounding boxes of glyphs. Therefore, it allows to query visual elements at different levels of semantic structure, i.e.,

page, t'ol, glyph-block, individual glyph, etc. The repository will also allow to query and study statistics of the Mayan writing system, e.g., hieroglyph co-occurrences.

Fig. 7: Snapshot of the online tool that feeds the database with imagery data and its corresponding annotations, i.e., codex name, t'ol, glyph-block reference, Thompson and Macri and Looper catalogs.

The second research line is the advancement of visualization techniques, and more precisely, the development of techniques that will allow exploring the feature space of a number of visual shape descriptors used to represent Mayan hieroglyphs for retrieval purposes. By relying on these visualization methods, our goals are detecting, understanding, and interactively overcoming some of the drawbacks associated with the shape descriptors currently in use (Vondrick, 2013).

5. Conclusions

We presented an overview of the MAAYA project's work-in-progress on epigraphic analysis, automatic visual analysis, and data access and visualization. Our close integration of work in computing and epigraphy is producing initial steps towards the design of computing methods tailored for epigraphy work; and can create opportunities to revisit findings in Maya epigraphy under the light of what computer-based methods can reveal (e.g., data-driven analyses of glyph diagnostic features.) At the same time, several of our machine learning, computer vision, and information retrieval methods are applicable to other problems in digital humanities. We would be interested in investigating applications of these methodologies to other sources of Cultural Heritage materials.

Acknowledgments.

We thank the support of the Swiss National Science Foundation (SNSF) and the German Research Foundation (DFG). We also thank all the members of the team (Rui Hu, Gulcan Can, April Morton, Oscar Dabrowski, and Peter Biro) for their contribution.

References

- CodeX Dresden (1962). *Codex Dresdensis: Maya Handschrift der Sächsischen Landesbibliothek Dresden*. Edited by the Sächsische Landesbibliothek Dresden from Prof. Dr. phil. habil. Eva Lips. Akademie-Verlag GmbH. Berlin. (Issue 626 from 700 printed.)

Codex Dresden (1989). *Die Dresdner Maya-Handschrift. Sonderausgabe des Kommentarbandes zur vollständigen Faksimile-Ausgabe des Codex Dresdensis.* Akademische Druckerei- und Verlags-Anstalt, Graz 1989, including Helmut Deckert: Zur Geschichte der Dresdner Maya-Handschrift and Ferdinand Anders: Die Dresdner Maya-Handschrift.

Codex Madrid (1967). *Codex Tro-Cortesianus.* Museo de América Madrid. Facsimilar Edition 1967 Moderated by Francisco Sauer and Josepho Stummvoll. Introduction and Summary by F. Anders. Akademische Druck- und Verlaganstalt Graz- Austria.

Evrenov, E.B., Kosarev, Y. and Ustinov, B.A. (1961). *The Application of Electronic Computers in Research of the Ancient Maya Writing.* USSR, Novosibirsk.

Förstemann, E.W (1880). *Die Maya-Handschrift der königlichen Bibliothek zu Dresden,* hrsg. von Ernst Wilhelm Förstemann. - Leipzig : Verlag der Naumann'schen Lichdruckerei, 1880.

Macri, M. and Vail, G. (2008). *The New Catalog of Maya Hieroglyphs, Volume Two: The Codical Texts.* University of Oklahoma Press, 308 pp.

Ngjam, J. (2011). *Unsupervised feature learning and deep learning tutorial.* http://ufldl.stanford.edu/wiki/index.php/UFLDL_Tutorial.

Roman-Rangel, E., Pallas, C., Odobez, J.-M. and Gatica-Perez, D. (2011). *Analyzing Ancient Maya Glyph Collections with Contextual Shape Descriptors,* Int. Journal of Computer Vision, Special Issue on e-Heritage, Vol. 94, No. 1, pp. 101-117, Aug. 2011.

Roman-Rangel, E. and Marchand-Maillet, S. (2013). *Stopwords Detection in Bag-of-Visual-Words: The Case of Retrieving Maya Hieroglyphs.* International Workshop on Multimedia for Cultural Heritage (MM4CH), at International Conference on Image Analysis and Processing.

Roman-Rangel, E. and Marchand-Maillet, S. (2013). *Bag-of-Visual-Phrases via Local Contexts.* Workshop on Recent Advances in Computer Vision and Pattern Recognition (RACVPR), at Asian Conference on Pattern Recognition.

Thompson, J. E. S. (1964). *A Catalog of Maya Hieroglyphs.* University of Oklahoma Press. Available online at: <http://www.famsi.org/mayawriting/thompson/index.html>

Villacorta, J. A. and Villacortax, C. A. (1976) *Códices Mayas (reproducidos y desarrollados por).* Sociedad de Geografía e Historia de Guatemala, Guatemala, C.A. (second edition.)

Vondrick, C., Khosla, A., Malisiewicz, T., Torralba, A. (2013). *HOGgles: Visualizing Object Detection Features.* International Conference on Computer Vision.

Zimmermann, G. (1956). *Die Hieroglyphen der Maya Handschriften.* Abhandlungen aus dem Gebiet der Auslandskunde, Band 62- Reihe B, Universität Hamburg.

SLUB: Sächsischen Landes- und Universitätsbibliothek Dresden
<http://digital.slub-dresden.de/werkansicht/dlf/2967/1/cache.off>
<http://gallica.bnf.fr/ark:/12148/btv1b8446947/f1.zoom=r:Codex%20Peresianus.langDE>
http://digital.slub-dresden.de/werkansicht/cache.off?id=5363&tx_dlf%5Bid%5D=2967&tx_dlf%5Bpage%5D=47
http://digital.slub-dresden.de/werkansicht/cache.off?id=5363&tx_dlf%5Bid%5D=2967&tx_dlf%5Bpage%5D=47

Automating the Search for Cross-language Text Reuse

Gawley, James

University at Buffalo, United States of America

Forstall, Christopher

University at Buffalo, United States of America

Clark, Konnor

University at Buffalo, United States of America

Tesserae is a open-source, online tool for detecting allusions in Classical literature on an automated basis. Originally limited to Latin poetry, the corpus of texts available to Tesserae has recently expanded to include Greek poetry and drama. Word-level n-grams form the foundation of the existing detection algorithm: a standard search returns all instances wherein two words a phrase in a later text shares two words with a phrase in an earlier text. This method has been previously demonstrated to reliably capture intertextual parallels already noted by philologists and to identify significant, previously unrecorded intertexts.

The ability to detect allusions across the language barrier would represent an evolutionary expansion in Tesserae's functionality as well as a significant contribution to Classical philology. Roman poets openly acknowledged their indebtedness to Greek literature (Horace famously remarked, "Greece, being conquered, tamed its wild conqueror, and brought the Arts to rustic Latium") and scholarly studies of Latin poetry have long commented on allusions to earlier Greek sources. To apply the existing system where Latin text alludes to Greek, Tesserae requires a translation dictionary linking Greek lemmata to associated Latin lemmata. This paper details two methods for building such a dictionary on an automated basis and compares their relative merits as measured by their ability to capture parallels between book one of Vergil's *Aeneid* and the *Iliad* of Homer, as noted by G.N. Knauer in his commentary.

The first method represents an original application of Bayes' theorem to a word-by-word alignment of the Greek New Testament with Jerome's Latin Vulgate.

$$P(L_i|G_i) = \frac{P(G_i|L_i)P(L_i)}{P(G_i)}$$

For a given Greek word G_i , the set of Greek Bible verses in which it appears is identified. The words contained in the Latin translation of these verses become the set of possible translation candidates L . For each L_i , the set of possible Greek words G is gathered from the set of Greek verses corresponding to the Latin verses in which L_i appears. $P(G_i|L_i)$ is represented by the number of words in set G which may share a lemma with G_i , divided by the total number of words in that set. The probability of G_i is represented by a similar calculation, where the set of all words within the Greek text is substituted for G . The value of $P(L_i)$ is analogous. The success of this relatively simple alignment algorithm as compared with more classical IBM Models or Hidden Markov Models may be explained by the grammatical similarity of these two inflected languages and importance placed by the translator in remaining precisely faithful to the syntax of the original text.

The second method employs English as a pivot language, in a method inspired by work done previously by Jeffrey Rydberg-Cox at Perseus on Latin-Greek synonymy. Using the XML-encoded digital editions of Lewis and Short's Latin-English Lexicon and Liddell and Scott's Greek-English Lexicon, two dictionaries widely considered authoritative for Classical languages and available through the Perseus Digital Library, each Latin or Greek headword is characterized by a feature set composed of the English words appearing in its definition. The Python-based Gensim topic modelling tools are then used to transform the English word counts to TF-IDF weights and calculate similarities between the dictionary entries. The similarity scores between entries are then interpreted as similarities in meaning between the respective headwords.

Each of the two methods described above produces pairwise similarities between all Greek and Latin words considered, with those pairings rated by a probability measure between 0 and 1. Because each Greek word may have more than one possible Latin translation, each method accepts the top two translation candidates as valid.

The text of Homer's *Iliad* is then indexed according to a feature set made up of Latin translation candidates. Each Greek token is lemmatized, and the token is then indexed according to all possible Latin translation candidates. Because lemmatization is unsupervised, ambiguous forms may have multiple possible Greek lemmata. Each possible Greek lemma will have two

translation candidates if the respective translation method is successful, or zero if no translations are found. The text of Vergil's Aeneid is indexed simply according to the possible Latin lemmata of each token. A given token in Vergil matches a token in Homer where one or more possible lemmata for the Latin word match against the set of translation candidates for the Greek word. A pair of phrases, one in Greek and the other in Latin, which share two or more words that match in this way, is returned as a possible allusion.

The two methods are evaluated by their ability to detect a subset of *Aeneid-Iliad* parallels collated from the commentary of G.N. Knauer. Each method retrieves a distinct, though partially overlapping, subset of the parallels noted by Knauer. Comparison of the respective performance of both methods suggests that, while each method can be shown to identify significant Latin-Greek allusions, the Bayesian alignment method provides better recall of the benchmark set than the 'pivot' method at the expense of precision. We ultimately aim to combine the output of both approaches into a single feature set.

References

- Tesserae, tesserae.caset.buffalo.edu (Accessed on November 1, 2013).
- Coffee, N. et. al** (2012).: "Intertextuality in the Digital Age." Transactions of the American Philological Association, Volume 142, Number 2, Autumn 2012 pp. 383-422
- Epistles*, 2.1.156–7
- G.N. Knauer** (1964): "Die Aeneis und Homer: Studien zur poetischen Technik Vergils mit Listen der Homerzitate in der Aeneis." Gottingen: Vandenhoeck & Ruprecht.
- Personal communication with author; tool archived at perseus.mpiwg-berlin.mpg.de/PR/syn.ann.html
www.perseus.tufts.edu
radimrehurek.com/gensim

XML-Print. Typesetting arbitrary XML documents in high quality

Georgieff, Lukas
lukas.georgieff@hotmail.com
UAS Worms

Küster, Marc Wilhelm
kuester@fh-worms.de
UAS Worms

Selig, Thomas
selig@fh-worms.de
UAS Worms

Sievers, Martin
sievers@uni-trier.de
University of Trier

Introduction/Motivation

Also in the age of electronic publishing print publications often remain the points of reference. While many humanities' projects finally build on XML and in particular TEI¹ and embrace electronic publications, they still want or need to target print publications as one or even the main form of sharing the results of their scholarship with the community. Paper remains the principal scholarly format accepted in many circles and in spite of all activities on long-term digital archiving^{2 3}, it remains a central medium to disseminate and conserve patrimonial content over the decades and centuries. However, how can you combine an XML-based workflow with the need to publish high-quality, multilingual print output respecting the often arcane typesetting requirements of scholarly texts in humanities publishing and notably in the realm of critical editions? While there are commercial^{4 5} and free^{6 7 8} products out there that

can help for some parts of the job, they are too expensive or too difficult to use in most humanities projects. This was the starting point of the DFG-funded XML-Print, an Open Source project that tackles the typesetting requirements for multilingual critical editions while offering a user-friendly frontend. XML-Print has already been presented to the DH community in two well-received short papers^{9 10} on specific aspects of the project's progress and technical challenges. In this long paper we present the project's overall results.

Typesetting Features

XML-Print consists of two components:

- an interactive graphical user interface (GUI) based on Eclipse, to define the rules for typesetting the XML texts in direct interaction with the source XML text
- a command-line typesetting engine, written from bottom up in F#, to actually transform the XML text into pdf for print

Normally scholars will interact with the GUI to map their XML structures on layout rules. The typesetting engine is then transparent to them. However, more automated workflows can integrate the engine directly. In section 5 we present examples for both scenarios. Beyond most standard typesetting functionalities of XSL-FO^{11 12} XML-Print supports in particular the following three requirements specific for publishing in humanities scholarship:

Columns

A page always consists of different rectangular regions to add header/footer, marginalia and the main text. However, this main text is often not limited to a one-column layout, but rather flows in multiple columns. As XSL-FO lacks in this requirement, XSL-FO+ adds a special interface to set-up arbitrary complex column-layouts, even mixed on one page. Each column has its own width and writing direction (left-to-right vs. right-to-left) allowing even "exotic" layouts to be applied within XML-Print.

Cross references

When using cross-references we must use placeholders, not only in the main text, but also for the header and footer of a page, where page number and sectioning information are commonly used, and for apparatus' entries, where typographic information like referenced line numbers can change during the editorial process.

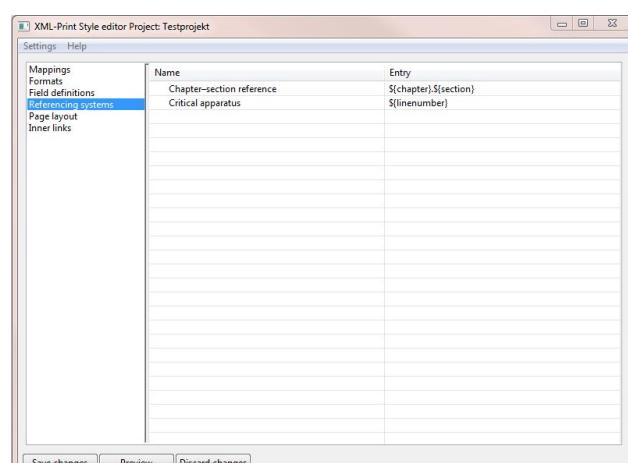


Fig. 1: Two user-defined reference systems for an edition

XML-Print incorporates a concept of "reference fields" to define structural and typographic elements to be counted. This way the user can even combine these two types, e.g. for having a global line count and a local one being reset at a specific XML structure. In addition the corresponding "title" of a reference field can as well be made available.

Apparatuses

Based on reference fields users can define “referencing schemas” to be used in apparatuses. Predefined typographic elements like a global page, column and line numbering can be combined with user-defined fields, arbitrary fixed strings and special characters, e.g. a non-breakable space. As the concrete output of the schema might depend on previous apparatus’ entries, exceptions with regard to repeated items, e.g. same chapter, can be formulated as well.

General Architecture and Technologies

Standards

Modern functional programming languages reusing established frameworks and libraries allow to build a high-quality, multilingual typesetting engine generating archiving-ready PDF/A-1¹³ much faster than even a decade ago. With a comparatively small development team we have been able to meet the project’s major objectives within the funding period.

XML-Print builds on Open Standards, especially on XML as the input language, XSL-FO to express formatting and XSLT to transform data from XML to XSL-FO. The project has extended XSL-FO to cover features such as apparatus and advanced referencing not currently supported by the specification (XSL-FO+). For the typesetting engine the project uses the .NET functional language F#¹⁴, running on the cross-platform Open Source .NET implementation mono. To handle OpenType¹⁵ fonts and generate pdf we have settled on the Open Source library iText¹⁶ that exists for both .NET and Java. We contributed to the library’s support for some of advanced OpenType features such as “real” small caps and aspects of bidirectional scripts.

Algorithms

A major advantage of functional programming languages is the lack of mutable variables and states. Algorithms are commonly more compact and easier to parallelize without mutable variables to share across multiple threads. The XML-Print backend for example parallelizes the parsing of certain XSL-FO elements and the rendering of chapters respectively page sequences.

Initially the rendering module was mainly based on the iText library. Now we are replacing all iText algorithms by our self-developed algorithms. They are specialized on the requirements of XML-Print and so are more efficient and easier to extend. We also decided to develop an own line-breaking algorithm. Going beyond the algorithm by Knuth and Plass¹⁷, we take advantage of today’s hardware capacities.

The line breaking algorithm creates a tree structure for all possible line combinations of an entire paragraph. The best path in this tree structure is calculated by taking several characteristics into account, e.g. interword spaces, hyphenations, etc. The final implementation will be parallelized and produce a tree structure with “cross-connected” nodes, i.e. nodes that represent identical paragraph sections are reduced and replaced by additional arcs, thus increasing the efficiency by avoiding redundant line combinations. Figure 1 illustrates the process of line breaking. Each node represents a possible line. The numbers at the arcs represent the processing order. Equal numbers on the same level mean a parallelized section. The bolded path represents the final paragraph, consisting of the nodes Line_1^2, Line_2^4^5, Line_3^2 and Line_4^4.

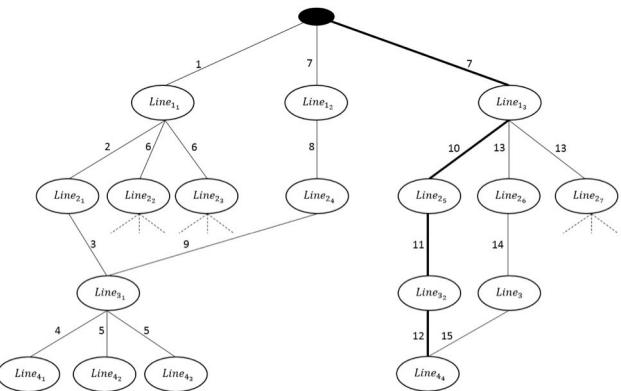


Fig. 2: Line breaking algorithm

Graphical User Interface

XML-Print addresses beginners as well as expert users. For the latter the GUI has to offer enough details while the former should not be overwhelmed by too many information at first. To achieve this goal XML-Print categorizes functionality and provides a basic and an expert layer.

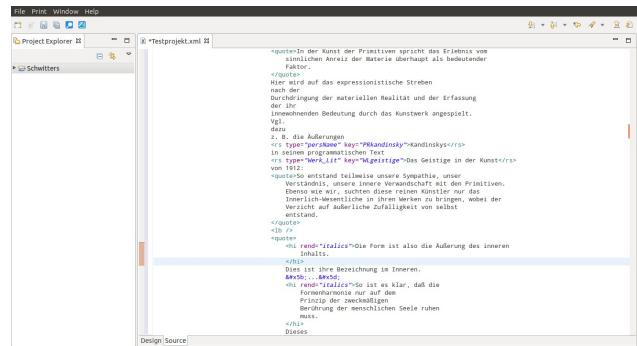


Fig. 3: GUI for XML-Print

We face, however, one inherent problem. To guarantee a high-quality output the typesetting incorporates a rich repertoire of typesetting logic and features which the user expects to appear somewhere in the graphical user interface. It is not always possible to shield users from these inherent complexities, while mapping all possible options onto GUI elements.

Apart from defining formats and declaring "mappings" between XML elements and corresponding formats, the GUI offers several possibilities to modify aspects of typesetting, from preprocessing the XML source to altering the PDF output format, from configuring the page size to influencing hyphenation.

Use Case Examples

Edition “Kurt Schwitters: Wie Kritik zu Kunst wird”

During the starting phase of the XML-Print project, staff members of the editorial project “Kurt Schwitters: Wie Kritik zu Kunst wird”¹⁸ already used recent version of the software to proofread their XML transcriptions. At a later stage, formats and mappings for critical and commentary apparatus were added.

Dictionaries: The “Trierer Wörterbuchnetz”

The “Deutsche Wörterbuch von Jacob und Wilhelm Grimm” is a digitized version of the leading German Dictionary with more than 300.000 entries stored as XML inside a database. The pdf is a byproduct, creating pdf files on the fly is the only effective approach. The XML-Print typesetting engine was wrapped via a simple webservice interface, allowing remote access.

Decoupling typesetting engine and GUI improves on flexibility and scalability, as it adds the options of cluster-processing and batch processing to the standard interactive processing.

Outlook

To ensure the long-term viability of the project beyond the end of funding in May 2014 XML-Print integrates with TextGrid¹⁹ to complement the current stand-alone mode. In parallel we build up a community on SourceForge <http://sourceforge.net/projects/xml-print/>, reaching even now more than 50 monthly downloads on average. Further development of XML-Print is also intimately linked to the needs of its customers, especially the critical editions using it, evolving in response to specific requirements. XML-Print is there to play a key role in creating, sharing and preserving our digital and non-digital textual cultural heritage and humanities digital resources on one of the most durable media yet known to humankind – paper.

References

1. **Burnard, Lou, and Syd Bauman.** (2007). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Text Encoding Initiative.
2. **Hedges, M, A Jordanous, S Dunn, C Roueche, Marc Wilhelm Küster, Thomas Selig, M Bittorf, and W Artes.** (2012). *New Models for Collaborative Textual Scholarship*. In 6th IEEE International Conference on Digital Ecosystems Technologies (DEST), 1–6. doi:10.1109/DEST.2012.6227933.
3. **Neuroth, Heike.** (2009). Nestor Handbuch.
4. Global Publishing Solution. (2013). 3B2 (Aka APP, Arbortext Print Publisher). 2013 ed. Accessed November 29. www.3b2.com/3b2/.
5. Adobe- 2013. Desktop Publishing (DTP), Digital Publishing | Adobe InDesign CC, www.adobe.com/lu_de/products/indesign.html , accessed: 29-Oct-2013.
6. **Knuth, D E.** 1986. *The TEXbook (Computers & Typesetting Volume a)*.
7. **Apache Foundation.** (2013). *Apache FOP - a Print Formatter Driven by XSL Formatting Objects and an Output Independent Formatter*. [xmlgraphics.apache.org/fop/](http://xmlgraphics.apache.org/xmgraphics.apache.org/fop/).
8. **Lamport, Leslie.** (1986). *LaTeX: User's Guide & Reference Manual*. Addison-Wesley 1986.
9. **Burch, Thomas, Martin Sievers, Marc Wilhelm Küster, Claudine Moulin, Roland Schwarz, and Yu Gan.** (2012). *XML-Print: an Ergonomic Typesetting System for Complex Text Structures*. In Abstracts of Digital Humanities 2012, 375–379. Hamburg. www.dh2012.uni-hamburg.de/wp-content/uploads/2012/07/HamburgUP_dh2012_BoA.pdf.
10. **Küster, Marc Wilhelm, Thomas Selig, Lukas Georgieff, Martin Sievers, and M Bittorf.** (2013). *XML-Print: Addressing Challenges for Scholarly Typesetting*. In Abstracts for Digital Humanities 2013, 269–272. Lincoln, NE. dh2013.unl.edu/abstracts/files/downloads/DH2013_conference_abstracts_print.pdf.
11. **Berglund, Anders**, ed. (2006). *Extensible Stylesheet Language (XSL) Version 1.1*. W3C. www.w3.org/TR/2006/REC-xsl11-20061205/.
12. **Lawson, D.** (2002). *XSL-FO: Making XML Look Good in Print*.
13. ISO, and ISO TC 171 SC 2. (2005). *ISO 19005-1:2005 Document Management -- Electronic Document File Format for Long-Term Preservation -- Part 1: Use of PDF 1.4 (PDF/a-1).* 19005. 1st ed. ISO.
14. **Smith, Chris.** 82012). *Programming F# 3.0*. 2nd ed. O'Reilly Media. www.amazon.de/Programming-F-3-0-Chris-Smith/dp/1449320295.
15. **Microsoft, Adobe.** 2009. “Microsoft Typography - OpenType Specification.” Microsoft.com. 21.09.2009. www.microsoft.com/typography/otspec/default.htm. Accessed 2013-10-29.
16. “iTText Core | iText Software.” 2013. “iTText Core | iText Software.” [Itextpdf.com](http://itextpdf.com). itextpdf.com/product/itext. Accessed: 2013-10-29.
17. Stanford University. **Computer Science Dept, D E Knuth, and M F Plass.** (1980). *Breaking Paragraphs Into Lines*.
18. **Kocher, Ursula, and Isabel Schultz.** (2013). *Wie Kritik Zu Kunst Wird: Kurt Schwitters' Strategien Der Produktiven Rezeption*. Avl.Uni-Wuppertal.De. www.avl.uni-wuppertal.de/forschung/projekte/wie-kritik-zu-kunst-wird.html.
19. **TextGrid Konsortium.** (2010). *TextGrid: Über TextGrid*. www.textgrid.de/ueber-textgrid.html.

Let DH Be Sociological! [Short Paper]

Goldstone, Andrew
Rutgers University, US

Diagnosis

This paper is a diagnosis and a polemic. It takes as its occasion the startling recent popularity of topic modeling among practitioners of the digital humanities (Nelson, 2010; Weingart and Meeks, 2012; Jockers, 2013; Tangherlini and Leonard, 2013; Laudun and Goodwin, 2013). As diagnosis, I propose that the significance of topic modeling can be contextualized within the rising predominance of social, political, and cultural themes as the major interests of literary scholarship in the last forty years. This predominance can, I show, itself be concretely grasped using topic modeling, as in the three figures below.

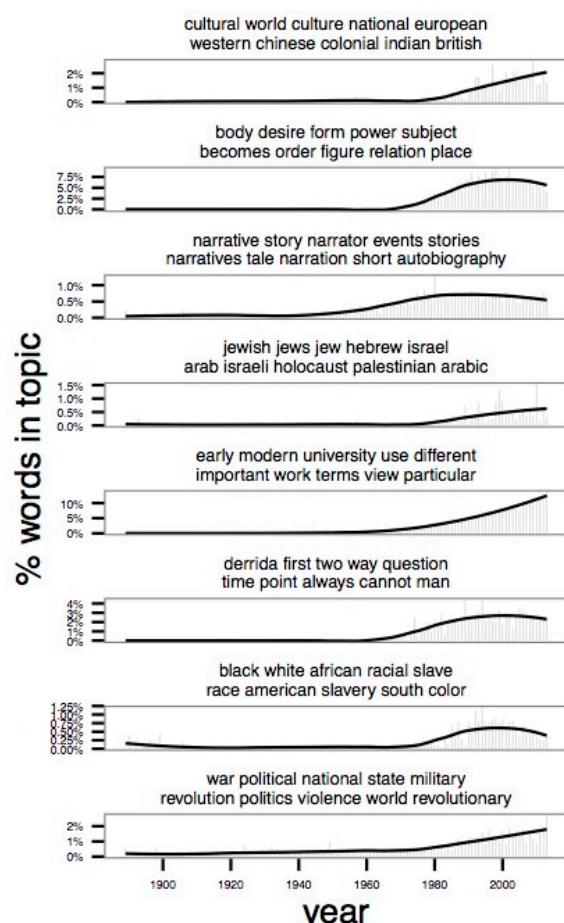


Fig. 1: Yearly proportions of recently-rising topics in a model of seven literary studies journals, labeled by most frequent words. Continued in figure 2.

These figures visualize all the recently-rising topics in a 120-topic model of a corpus of literary-studies articles from seven generalist journals from the 1889–2013 period. Before looking at time series, I coded each topic as “social/political,” “formal,” “other themes,” or “non-thematic.” Of the 26 recently-rising topics shown in the figures, 14 are classifiable as “social/political”; of the remaining 94, only 4 are. (See “Methods” for details). I argue that the recent turn in the digital humanities to computational studies of literary and cultural texts in the aggregate, typified by the work of Franco Moretti and Matthew Jockers (Moretti, 2013; Jockers, 2013), is best understood as an incomplete methodological response to an already-existing dominant thematic trend in literary studies.

Polemic

This historical diagnosis then leads to my polemic: let digital humanities be sociological! Instead of insisting on the distinctiveness of a “humanistic” interpretive approach—as, for example, Alan Liu has recently done in a sharp critique of “tabula rasa” digital interpretation—humanists should recognize the problem of interpreting cultural text in the aggregate as one they share with social science (Liu, 2013). This recognition can, in turn, help to clarify the controversy over whether the digital humanities deliberately neglect the social and political concerns central to literary and cultural studies in the last four decades (McPherson, 2012). Recognizing the sociological in the digital humanities would help to see how quantitative methods could address the fundamental concerns that humanists share with social scientists.

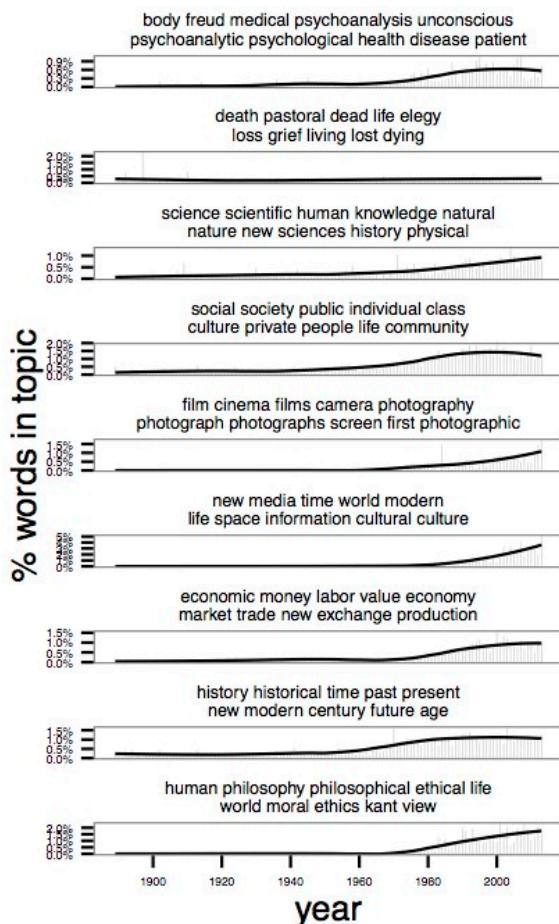


Fig. 2: Continued from figure 1; continues in figure 3.

In this short paper, I focus on the case of topic modeling: though this technique emerges from machine learning (Blei et al., 2003) and has been discussed as a form of distant reading, I argue that topic-modeling analyses of literary material

(including my own in this paper) should be categorized as *content analyses* in the social-scientific sense. Although connections between content analysis and humanities computing are of long standing (see Weber, 1985), the relevance of this methodology for topic modeling has not been widely remarked in the digital humanities. According to a standard book on the technique, “Content analysis is a research technique for making replicable and valid inferences from texts...to the contexts of their use” (Krippendorff, 2013, p. 24). The triple demands for validity, context-sensitivity, and replicability represent the fundamental social-scientific methodological contribution to this work. In work on topic modeling, these methodological problems have been addressed especially by political scientists (Quinn et al., 2010; Grimmer, 2010; Grimmer and Stewart, 2013) and sociologists of culture (DiMaggio et al., 2013; for a recent work on validation with literary topic models, see Mimmo and Jockers, 2013).

Topic modeling should not be valued as a tool for discovery alone but as offering evidence of systematic cultural variation. Emphasizing discovery (e.g., Blei, 2012), has led some to insist that the final task for humanistic topic modelers should be to return to “close reading” individual texts (Rhody, 2012; Tangherlini and Leonard, 2013). Yet this return to reading risks neglecting both the promise and the challenge of the topic model, which can reveal the workings of larger-scale cultural and social contexts by systematically and replicably classifying linguistic patterns, including thematic and rhetorical patterns.

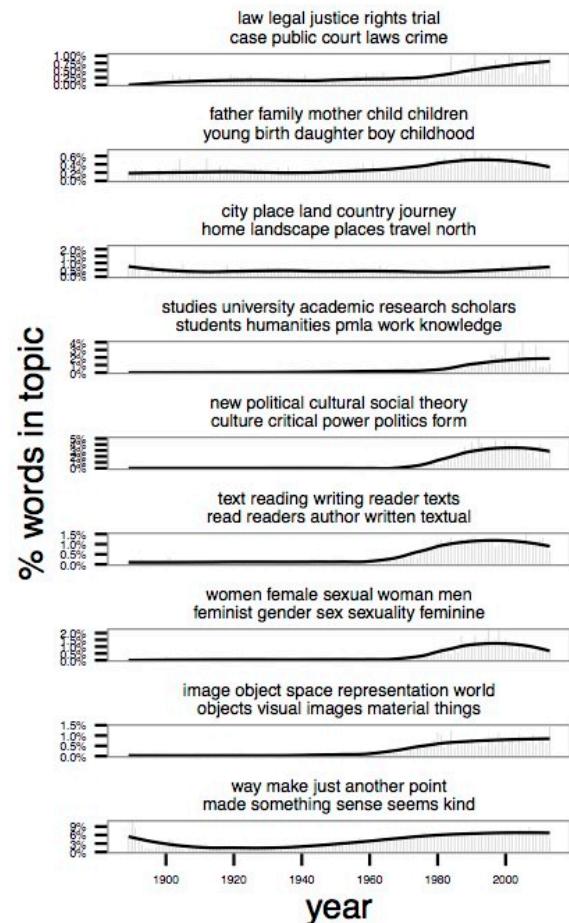


Fig. 3: Continued from figure 2

These patterns are of interest not in themselves alone but for their cultural, historical, and social contexts.

In my own argument, the category of “recent decades,” which highlights the rise of “social” topics, is actually a proxy for historical causes, including the institutional change represented in the corpus by the inclusion of “theory” journals newly established in the 1970s, Critical Inquiry and New Literary

History; it remains for future work to incorporate indicators of these historical forces into the analysis of topic models.

Even this preliminary content analysis suggests that digital humanists who study texts in the aggregate might reconsider the context in which their own work emerges. Current discussions of “distant reading,” “macroanalysis,” “surface reading,” and “quantitative formalism” converge with sociology in terms of method but not necessarily subject matter (Moretti, 2013; Jockers, 2013; Best and Marcus, 2009; Allison et al., 2010). At the same time, the aggregate of literary studies has been converging with sociology in terms of subject matter but not method, and even the most recent turns to the sociological in literary studies have largely shied away from the quantitative approaches that digital humanists have embraced (see English, 2010). My polemical goal is to advocate for a dual convergence—not only in the case of topic modeling but across the set of quantitative techniques for studying cultural texts that have become central to the digital humanities.

Methods

Latent Dirichlet Allocation has been applied to corpora of scholarly journals by others (Blei and Lafferty, 2009; Mimno, 2012; McFarland et al., 2013); this work applies it to scholarship in literary studies, with the institutional history of the literary humanities as an interpretive frame. The modeled documents consisted of all the items classed as “full-length articles” by JSTOR that exceed 1000 words in length in seven journals chosen for chronological range and broad disciplinary scope: *Critical Inquiry* (1974–2013), *ELH* (1934–2013), *Modern Language Review* (1905–2013), *Modern Philology* (1903–2013), *New Literary History* (1969–2012), *PMLA* (1889–2007), and *the Review of English Studies* (1925–2012). Wordcounts and document metadata were supplied by JSTOR Data for Research (JSTOR).

Obvious item misclassifications were corrected. I excluded an extensive set of stop words, including common words, abbreviations, and first names, and retained only the 10000 most frequent word types. MALLET’s Latent Dirichlet Allocation implementation was used, specifying 120 topics and hyperparameter optimization feature (McCallum, 2002). The choice of documents to model and the construction of the stoplist emerged from work by Ted Underwood and me; Underwood should not be held accountable for this paper (Goldstone and Underwood, 2012; Goldstone and Underwood, forthcoming). Additional analysis relied on the R mallet package (Mimno, 2013) and my own R programs.

The procedure for classifying the topics was as follows. I conducted a trial run by hand-classifying a 64-topic model of *PMLA* articles alone, developing an ad hoc scheme. Then, before visualizing any topics over time in my full seven-journal model, I examined the list of the twenty most frequent words in each topic and applied the categorization scheme to each topic:

S: Social or political topics, including national, ethnic, sexual, or gender identities;

T: Other thematic material, including religion, moral philosophy, love, nature, etc.;

F: Formal topics, including form, language, style, and genre;

NT: Non-thematic topics, including other languages, proper names, organizational labels, topics classifying textual studies, and clearly methodological discourses.

I classed as “recently rising” topics any topic for which the total proportion of those topics in each of the four decades after 1970 was greater than the total proportions in each of the decades from the 1930s through the 1960s. This heuristic was again devised with respect to the smaller trial model, then applied to the larger model. In future work ahead of the Lausanne conference, I plan to systematically vary the “recency” cutoff in order to test the sensitivity of my claims to the choice of 1970 as a demarcation line.

The breakdown of all topics was as follows:

code	not recent	recent
F	13	3
NT	56	4
S	4	14
T	21	5

By this effort to make interpretive assumptions explicit (and to highlight the involvement of the researcher in classifying topics), I seek to bring the humanistic analysis of topic models closer to the demands of sociological content analysis.

References

- JSTOR. Data for Research. <http://dfr.jstor.org>.
- Allison, S., et al. (2011). *Quantitative Formalism: An Experiment*. Stanford Literary Lab. <http://litlab.stanford.edu/LiteraryLabPamphlet1.pdf>.
- Best, S., and S. Marcus (2009). *Surface Reading: An Introduction*. Representations 108(1).
- Blei, D. M. (2012). *Probabilistic Topic Models*. Communications of the ACM 55(4).
- Blei, D. M., A. Y. Ng, and M. I. Jordan (2003). *Latent Dirichlet Allocation*. Journal of Machine Learning Research 3.
- Blei, D. M., and J. D. Lafferty (2009). *Topic Models*. In Srivastava, A., and M. Sahami (eds.). *Text Mining: Classification, Clustering, and Applications*. Boca Raton, FL: CRC.
- DiMaggio, P., M. Nag, and D. Blei (2013). *Exploiting Affinities between Topic Modeling and the Sociological Perspective on Culture: Application to Newspaper Coverage of U.S. Government Arts Funding*. Poetics 41(6).
- English, J. F. (2010). *Everywhere and Nowhere: The Sociology of Literature After “The Sociology of Literature.”* NLH 41(2).
- Goldstone, A., and T. Underwood (2012). *What Can Topic Models of PMLA Teach Us About the History of Literary Scholarship?* Journal of Digital Humanities 2(1).
- Goldstone, A., and T. Underwood (forthcoming). *The Quiet Transformations of Literary Studies: What Thirteen Thousand Scholars Could Tell Us*. NLH.
- Grimmer, J. (2010). *A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases*. Political Analysis 18(1).
- Grimmer, J., and B. M. Stewart (2013). *Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts*. Political Analysis 21(3).
- Jockers, M. L. (2013). *Macroanalysis: Digital Methods and Literary History*. Urbana: University of Illinois Press.
- Jockers, M. L., and Mimno, D. (2013). *Significant Themes in 19th-Century Literature*. Poetics 41(6).
- Krippendorff, K. (2013). *Content Analysis: An Introduction to Its Methodology*. 3rd ed. Los Angeles: SAGE.
- Laudun, J., and J. Goodwin (2013). *Computing Folklore Studies: Mapping over a Century of Scholarly Production through Topics*. Journal of American Folklore 126(502).
- Liu, A. (2013). *The Meaning of the Digital Humanities*. PMLA 128(2).
- 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).
- McPherson, T. (2012). *Why Are the Digital Humanities So White? or Thinking the Histories of Race and Computation*. In M. K. Gold (ed.), *Debates in the Digital Humanities*. Minneapolis: University of Minnesota Press.
- Mimno, D. (2012). *Computational Historiography: Data Mining in a Century of Classics Journals*. ACM Journal on Computing and Cultural Heritage 5(1).
- Mimno, D. (2013). *mallet: A Wrapper around the Java Machine Learning Tool MALLET*. <http://cran.r-project.org/web/packages/mallet/>.
- Moretti, F. (2013). *Distant Reading*. London: Verso.
- Nelson, R.K. (2010). *Mining the Dispatch*. <http://dsl.richmond.edu/dispatch>.

- Quinn, K. M., et al.** (2010). *How to Analyze Political Attention with Minimal Assumptions and Costs*. American Journal of Political Science 54(1).
- Rhody, L. M.** (2012). *Topic Modeling and Figurative Language*. Journal of Digital Humanities 2(1).
- Tangherlini, T. R., and P. Leonard** (2013). *Trawling in the Sea of the Great Unread: Sub-Corpus Topic Modeling and Humanities Research*. Poetics 41(6).
- Weber, R. P.** (1985). *Basic Content Analysis*. Beverly Hills, CA: SAGE.
- Weingart, S., and E. Meeks** (eds.) (2012). Special Issue: Topic Modeling. Journal of Digital Humanities 2(1).

Building a metrical ontology as a model to link digital poetic repertoires

González-Blanco, Elena

egonzalezblanco@flog.uned.es

Universidad Nacional de Educación Distancia, España

Seláf, Levente

levente.selaf@gmail.com

ELTE University, Budapest

Del Rio Riande, María Gimena

guineveregime@hotmail.com

Scerit-CONICET, Argentina

Martínez Cantón, Clara Isabel

cimartinez@flog.uned.es

Universidad Nacional de Educación Distancia, España

Martos Pérez, María Dolores

mdmartos@flog.uned.es

Universidad Nacional de Educación Distancia, España

1. Introduction

1.1. Overview

The technology of semantic web seems to be a suitable environment to offer solutions for linking poetic repertoires that belong to different European traditions and literatures (González-Blanco & Seláf 2013)¹. The problem of linking poetic repertoires is not simple, as there are not only technical issues involved, but also conceptual and terminological problems: each repertoire belongs to its own poetical tradition and each tradition has developed its own analytical terminology for years in a different and independent way. The result of this uncoordinated evolution is a bunch of varied terminologies to explain analogous metrical phenomena through the different poetic systems whose correspondences have been hardly studied.

1.2. Methodology

The aim of this paper is to present a model able to serve as a uniform solution for terminological issues in order to build a solid semantic structure as a basis to link the different poetic systems. This structure will be used to publish repertoires on the web in a structured format and using open standards in order to build an open-source and collaborative platform based on a poetic ontology which lets interoperability among the different European metrical repertoires with different applications, such as faceted searches based on SPARQL or different kinds of visualizations, very helpful for comparative analysis.

The first step to organize and manage repertoires and database systems was the construction of conceptual schema to define their basic entities and relationships. The ER (Entity-

Relationship) data model is the most commonly used for this purpose, together with the data model based on records for the logical implementation (Elmasri & Navathe 2011, 27-ss)², which is also widely accepted.

To implement this conceptual model, the project ReMetCa (Digital Repertoire on Medieval Spanish Poetry: www.uned.es/remetca) has tested different systems (commercial, free, open-source, and proprietary). The final decision, after experimenting with Oracle Express Edition (González-Blanco & Rodríguez 2013)³, has been MySQL combined with a XML tagging using the TEI-verse module. The relationship between ontological models and TEI is being taken into consideration very seriously in the last years, as it is shown by the activity of the SIG ontologies group wiki.tei-c.org/index.php/SIG:Ontologies and the specific papers published on this topic (Eide & Ore 2007)⁴. There are also projects that have applied these techniques to the study and analysis of medieval documents (Ciula, Spence & Vieira 2008)⁵.

2. Getting Started

From the three levels described (conceptual, logical and physical), this paper will focus on the first layer: the semantic description with the design of the semantic ontology, whose elements will be extensible and reusable for its application to other poetic repertoires. The conceptual model, designed on the basis of ReMetCa, will be transferred to the semantic Web as Linked Open Data. The abstraction of this initial model is prepared to be amplified with the necessary fields and terms to define metrical phenomena which are not shown in the Spanish poetic system or in the other repertoires which have been taken into account to design this first version of the semantic prototype. In order to enlarge its horizons, structure, description and contents, datasets of the following corpora have also been taken into account:

- The Cantigas de Santa María Database: csm.mml.ox.ac.uk
- Analecta Hymnica Digitalia: database on Medieval Latin poetry: webserver.erwin-rauner.de/crophius/
- Analecta_conspectus.htm
- Bibliografia Eletronica dei Trovatori: w3.uniroma1.it/bedt/BEdT_03_20
- Le Nouveau Naetebus: database on French narrative Medieval poetry : www.nouveunaetebus.elte.hu
- Répertoire de la Poésie Hongroise Ancienne (RPHA) : Repertoire on Medieval Hungarian poetry: rpha.elte.hu
- MedDB: Lírica Profana Galego-Portuguesa www.cirp.es/pls/bdo2/f?p=MEDDB2
- Corpus rhythmorum musicum (IV-IX secolo): database on Latin Medieval poetry accompanied with music www.corimu.unisi.it
- Skaldic poetry of Scandinavian Middle Ages: <https://www.abdn.ac.uk/skaldic/db.php>
- English Broadside Ballad: ebba.english.ucsb.edu/
- Dimev: Digital index of medieval English verse: www.cddc.vt.edu/host/imev/record.php?recID=6768

To implement the conceptual model, the project uses one of the most recognized standards for semantic Web description: the Ontology Web Language (OWL), developed by W3C as an extension of RDFS. OWL is used to define the different classes, their properties and the instances of classes. It integrates sets of predefined metadata using namespaces. The set joins not only traditional well-known initiatives, such as Dublin Core, MARC or TEI, but also local proposals such as those used by some of the digital poetic repertoires that serve as a basis for this project. The TEI-Verse module⁶ plays also an important role, due to the use that several repertoires have made of it, such as Henrik Ibsens (<http://www.ibsen.uio.no>) or the project of Lyrik des Hohen Mittelalters, (whose web access is not public yet), or ReMetCa itself by the addition of a XMLType field to its relational database.

Fig. 1: ReMetCa database screen with XML-type field

The software used to build the collaborative ontology is Webprotege (Tudorache et al. 2011)⁷, initially combined with Poolparty to create and organize vocabularies. It has been installed at ReMetCa server and opened via web in order to let participation of researchers with similar projects in the field of metrical repertoires. This system presents a light and intuitive interface but solid enough to develop a complex ontology with OWL. An important advantage is that it offers multilingual edition, which is very important for the development of such an international proposal. Once the model had been set, a metadata system has been designed to link the conceptual and logical levels based on a global abstract classification (schema), in which the different particular embodiments of each poetical tradition will be progressively included. This proposal shows both the consistency of this general language purpose and the benefits that can be obtained from the application of this model to the different local projects using a collaborative and open work system, which is essential for this new paradigm.

There are a few studies which deal obliquely with some of the above mentioned aspects (Bootz & Szonięcki 2008⁸ and Zöllner-Weber 2009)⁹, but there is not yet a conceptual model of ontology referred to metrics and poetry. The closest related works to this topic are probably the conceptual model of CIDOC (www.cidoc-crm.org), the vocabularies of the Getty Museum, as they are designed to express relations and artistic manifestations in the field of humanities (<http://www.getty.edu/research/tools/vocabularies/>), 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 a deep information on metrics vocabulary.

To sum up, this project of a poetic and metrical ontology intends to be much more than a repository of datasets, thesauri or controlled vocabularies. It aims to create a semantic standardized structure to describe, analyze and develop logical operations through the different poetic digital repertoires and their related resources. Its final objective is to interconnect, reuse and locate the data disseminated through poetic databases in order to get interoperability among projects, to perform complex searches and to make the different resources "talk" to each other following a unique but adaptive model.

References

1. **González-Blanco, E. & L. Seláf** (2013), *Megarep: A comprehensive research tool in Medieval and Renaissance poetic and metrical repertoires*, Humanitats a la xarxa: món medieval / Humanities on the web: the medieval world, edited by L. Soriano, M. Coderch, H. Rovira, G. Sabaté & X. Espluga., Oxford, Bern, Berlin, Bruxelles, Frankfurt am Main, New York, Wien, Peter Lang (forthcoming).

2. **Elmarsi, R. & S. B. Navathe** (2011), *Fundamentos de Sistemas de Bases de Datos*, Madrid, Pearson, Addison Wesley.

3. González-Blanco, E. & 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. Abstracts of the TEI Conference and Members Meeting 2013: October 2-5, Rome edited by Fabio Ciotti & Arianna Ciula, DIGILAB Sapienza University & TEI Consortium, Rome, 178-185. digilab2.let.uniroma1.it/teiconf2013/abstracts/. Accessed 30-10-2013.

4. Eide, Ø. & C.-E. Ore (2007), *From TEI to a CIDOC-CRM Conforming Model: Towards a Better Integration between Text Collections and Other Sources of Cultural Historical Documentation*, paper presented at the DH conference 2007. Abstract available at: www.edd.uio.no/artiklar/tekstkoding/poster_156_eide.html

5. Ciula, A., P. Spence & J. M. Vieira (2008), *Expressing complex associations in medieval historical documents: the Henry III Fine Rolls Project*, Literary and Linguist Computing, 23 (3): 311–325.

6. Burnard, L. & S. Bauman, eds., *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Ver. 2.5.0. Accessed 30-10-2013. www.tei-c.org/release/doc/tei-p5-doc/en/html/

7 Tudorache T C | Nyulas N E Nov & M A Musep

7. Tudorache, T., C. I. Nyulas, N.F. Noy & M.A. Musen (2011), *WebProtégé: A Collaborative Ontology Editor and Knowledge Acquisition Tool for the Web*, Semantic Web Journal, IOS Press. www.semantic-web-journal.net/content/webprot%C3%A9g%C3%A9-distributed-ontology-editor-and-knowledge-acquisition-tool-web. Accessed: 30/10/2013.

8. Bootz, P. & S. Szoniecky (2008), *Towards an ontology of the field of digital poetry*, paper presented at Electronic Literature in Europe. Full text available at elmcip.net/node/415

9. Zöllner-Weber, A., "Ontologies and Logic Reasoning as Tools in Humanities?", DHQ 2009, 3: 4. www.digitalhumanities.org/dhq/vol/3/4/000068/000068.html Accessed: 30/10/2013.

9. Zöllner-Weber, A. (2009), *Ontologies and Logic Reasoning as Tools in Humanities?*, DHQ, 3: 4. www.digitalhumanities.org/dhq/vol/3/4/000068/000068.html
Accessed: 30/10/2013.

Exploring Usage of Digital Newspaper Archives through Web Log Analysis: A Case Study of Welsh Newspapers Online

Gooding, Paul
paul.gooding.10@ucl.ac.uk
University College London

1. Introduction

This paper will present the findings from a case study undertaken with the support of the National Library of Wales. It utilises web log analysis to discover more about the use and users of Welsh Newspapers Online (WNO). [1] a digitised newspaper archive which currently contains around 420,000 digitised newspapers from, and relating to, Wales. Web log analysis has previously been undertaken to analyse online user behaviour, with utility for websites¹⁻², e-journals³⁻⁴⁻⁵, and digital resources⁶⁻⁷. To date, however, it has not been used to analyse the use of digitised newspaper archives. This paper presents the findings from a detailed analysis of content logs from a period of three months, starting from the launch of WNO in March 2014, and provides an important empirical study into how users are interacting with digitised newspapers. In doing so, it helps to illuminate some existing debates about the impact of digitisation upon reading and scholarship.

These debates have focused on how digitisation of newspapers has changed the way users engage with these materials. Mussell, for instance, has noted that article-level representation foregrounds the partial textual transcript, even though the physical article is actually one textual component

operating among many others. This means that users of a digitised newspaper may lose the original context of the material by approaching it in a different way online⁸. Brake additionally notes that "digital representations of nineteenth-century copy denaturalizes it and transforms the reader... into a user who sees the content inextricably embedded in the matrix of the newspaper pages"⁹. These issues have been expressed in concerns that this will have a sever impact on reading behaviour^{10 11 12}. This paper demonstrates that, while user engagement remains extremely high, it is evident that the nature of this engagement has been unavoidably transformed by digitisation.

2. Methodology

In order to undertake this analysis, the National Library of Wales IT team supplied a complete set of content logs, collected for a period of three months from the launch of the collection in March 2013. These logs specifically recorded interactions with content discover and viewing mechanisms on the WNO website. As a result, they are not a complete log of the user's journey. Instead, the logs track a number of important basic behaviours: searchers undertaken by users on the website (search queries); occasions where users have browsed, filtered, or otherwise interacted with search results (search results queries); and instances where users have viewed newspaper pages (content queries). The weblogs record each interaction with the website as a single line of plain text, recorded automatically on the website servers. The following example demonstrates the format of a single content query:

```
2013-06-02T12:26:50+01:00 51a5c97c3c8d3 llgc-id:3036868
llgc-id:3039814 llgc-id:3037695 Aberystwyth Observer 21
September 1872 [2] ART40
```

The elements are, in order: date and time of interaction; unique user ID; server IDS for website content; title of newspaper viewed; date of newspaper edition; page number viewed; article number on the page viewed. Additionally, search queries contain a field with the user's search terms, and search results queries include a field which shows the nature of the user's interaction. In total, there were over 300,000 weblogs, which therefore provide a rich source of information about user interactions with the content of WNO.

This data allows several metrics to be analysed: most viewed newspaper titles; most viewed years; most viewed page numbers; average number of pageviews per visit; and percentage of pageviews involving search, search results or content queries. To achieve this, the investigator heavily post-processed the data in Excel. All non-relevant fields were stripped from the spreadsheet, then a column was added which automatically populated pageview numbers for each unique user ID, from 1 for the first page viewed by a user, to X for the last. Search, browser and content queries were also changed to numerical values of 1,2, and 3 respectively. This allowed the data to be processed into appropriate graphical representations. In particular, it allowed large-scale analysis of the complete dataset, in order to uncover overall patterns and trends in graphical form.

3. Findings

The content logs have provided a wealth of fascinating data on the behaviour and interests of users of Welsh Newspapers Online. We will discuss some of the most important findings in more depth below.

The following chart shows the findings from the large-scale analysis of the content logs, demonstrating the percentage of queries by type at each pageview. From this chart we can see which broad categories of user behaviour were most common for any given pageview:

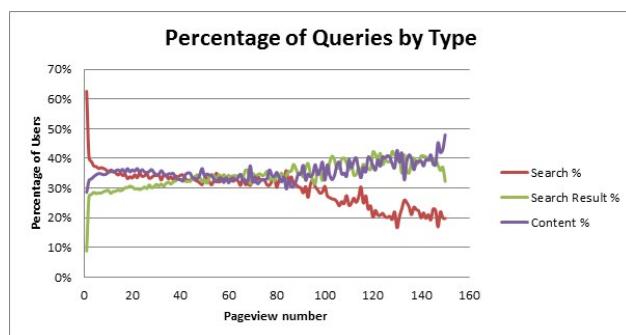


Fig. 1:

The chart demonstrates that content queries remain an important part of user activity, regardless of how long visitors spend on the website. In fact, once visitors view over 100 pages, they seem to view increasingly large amounts of content. By contrast, the longer a user spends on the resource, the less likely they are to be engaged in search activities. Instead, we see that search result queries replace search queries in importance.

This tells us that the importance of search as a discovery mechanism becomes less important with time spent on a particular visit: the majority of users begin by searching the collection, and then slowly but steadily move away from searching to view content or to browse search results. However, the data suggests that digital reproductions of newspapers do effect how users engage with them, in one particular way: while they browse the website, extensively, we can see that they do not browse through newspaper content. The following chart shows that users primarily tend to view the title page but, in general, the further into a newspaper the page is,

the fewer times it will be read. This data should be contextualised: the average number of pages per newspaper is around 6.3. However, the drop in views begins from the second page, which strongly suggests that users are unlikely to browse through pages:

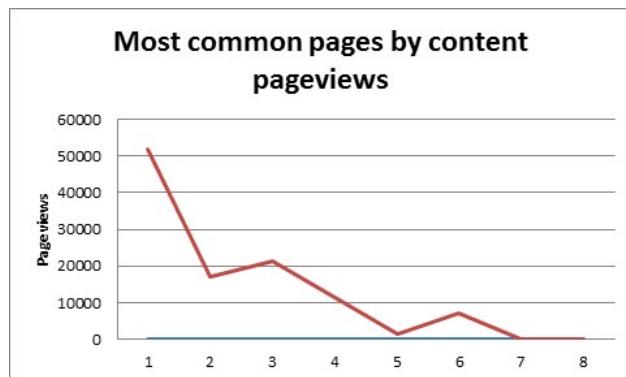


Fig. 2:

In this respect, the findings seem to confirm the fears that digitised texts interfere with the deep, ordered reading behaviour attributed to printed materials. While there are valid concerns about the way in which inline reading may distance users from the original context of newspaper material, we must also recognise that users certainly are deeply engaged with digitised newspaper archives: it is the type of engagement that has changed. This highlights two important conclusions: the side-lining of browsing makes serendipitous discoveries less likely, and means that search functionality, OCR quality, and website design are as important as the availability of content in ensuring that users are able to discover content; and that existing remediations of newspapers are some way from recreating the reading behaviour that has been attributed to the physical text^{13 14}.

4. Conclusion

This case study demonstrates that content log analysis can greatly enrich studies of the impact of digitised collections. They provide empirical evidence of user behaviour, which allows a granular, nuanced understanding of how researchers interact with digitised content online. We can therefore learn a great deal more about this understudied area. The findings also confirm, though, a commonly recorded problem with content logs; they provide a strong statistical base for analysis, but they cannot provide an insight into why users behave how they do. As such, an analysis which relies solely on content logs is necessarily incomplete. This paper will finish by proposing that web log analysis can still play an important role in analysing and understanding usage of digital resources, but that it should not be used in isolation.

[1] <http://papuraunewyddcymru.llgc.org.uk/en/home?>

References

1. Nicholas, D. et al., (2000). *Evaluating consumer website logs: a case study of The Times/The Sunday Times website*. Journal of Information Science, 26(6), pp.399–411.
2. Almind, T.C. & Ingwersen, P., (1997). *Infometric Analyses on the World Wide Web: Methodological Approaches to 'Webometrics'*. Journal of Documentation, 53(4).
3. Institute for the Future, (2002). *E-Journal user study: report of web log data mining*, Available at: ejust.stanford.edu/logdata.html [Accessed March 23, 2013].
4. Nicholas, D., Jamali, H.R. & Huntington, P., (2005). *The use and users of scholarly e-journals: a review of log analysis studies*. Aslib Proceedings, 57(6), pp.554–571.
5. Yu, L. & Apps, A., (2000). *Studying e-journal user behaviour using log files: the experience of SuperJournal*. Library and Information Science Research, 22(3), pp.311–338.
6. Warwick, C. et al., (2008). *If You Build It Will They Come? The LAIRAH Study: Quantifying the Use of Online Resources in the Arts and Humanities*. Literary and Linguistic Computing, 23(1), pp.85–102.
7. Meyer, E.T. et al., (2009). *Usage and Impact Study of JISC-Funded Phase 1 Digitisation Projects & the Toolkit for the Impact of Digitised Scholarly Resources (TIDSR)*, Oxford: Oxford Internet Institute, University of Oxford. Available at: microsites.oiil.ox.ac.uk/tidsr/sites/microsites.oiil.ox.ac.uk.tidsr/files/TIDSR_FinalReport_20July2009.pdf [Accessed May 9, 2013].
8. Mussell, J., (2012). *The Nineteenth-Century Press in the Digital Age*, Basingstoke: Palgrave MacMillan.
9. Brake, L.
10. Half Full and Half Empty. Journal of Victorian Culture, 17(2), pp.222–229.
11. Birkerts, S., (1994). *The Gutenberg elegies: the fate of reading in an electronic age*, New York: Ballantine Books.
12. Deegan, M. & Sutherland, K., (2009). *Transferred Illusions: Digital Technology and the Forms of Print*, Ashgate Publishing. p.49.
13. Baker, N., (2002). *Do we want to keep our newspapers?* In Do we want to keep our newspapers? London: Office for Humanities Computing, pp. 19–34.
14. Birkerts, S. (1994). *The Gutenberg Elegies: The Fate of Reading in an Electronic Age*. Boston: Faber and Faber.

DigCurV: curriculum framework for digital curation in the cultural heritage sector

Gow, Ann

ann.gow@glasgow.ac.uk
University of Glasgow

Molloy, Laura

laura.molloy@glasgow.ac.uk
University of Glasgow

Konstantelos, Leo

leo.konstantelos@unimelb.edu.au
University of Melbourne

1. Introduction

The Digital Curator Vocational Education ('DigCurV') project was funded by the European Commission's Leonardo da Vinci lifelong learning programme¹. It aimed to establish a curriculum framework for vocational training in digital curation.

DigCurV brought together a network of partners² to address the availability of vocational training for digital curators in the library, archive, museum and cultural heritage sectors, with a particular focus on the training needed to develop new skills that are essential for the long-term management of digital collections.

1.1. Overview

In 2013, the DigCurV collaborative network completed development of this Curriculum Framework for digital curation skills in the European cultural heritage sector.

Drawing on a variety of established skills and competence models in the digital curation and cultural heritage sectors, DigCurV synthesised such expertise with input from those in the digital curation professions to develop a new Curriculum Framework. As a result, the Framework can help develop digital curation training offerings, provide a benchmark against which to map and compare existing offerings, and motivate training providers to continue to develop and refresh training. As the range of digital humanities stretches across disciplines, such frameworks and lenses are essential for understanding the skills and competences of individuals and for describing roles.

Our paper will describe the salient points of this work, including how the project team conducted the research necessary to develop the Framework, the structure of the Framework, the processes used to validate the Framework, and three 'lenses' onto the Framework.

Our paper will also provide suggestions as to how the Framework might be used, including a description of potential audiences and purposes. As such, this paper draws on various DigCurV project deliverables. The contributions of members of the network to these deliverables is gratefully acknowledged.

1.2. Background

A critical and often sidelined issue within digital humanities, and the cultural heritage sector more widely, is the ability of those undertaking research in the arts and humanities to care for their data and other digital material over time. Digital humanities research creates rich digital resources³ but also the challenges of sustaining and managing these objects. Other professionals in the cultural heritage sector also have the responsibility of stewardship of digital material over time. But are those now professionally obliged to perform digital curation receiving the training they need? And what exactly constitutes those training needs?

Another pedagogical dilemma in digital curation is whether all staff in the digital humanities and cultural heritage sector should become more proficient in the curation of digital assets, or whether specific training should be developed to enable a distinct strain of specialists to emerge. As digital humanities scholars should we be skilled to care for as well as to create? The Italian economist Vilfredo Pareto argued at the turn of the twentieth century that a society grown wealthy enough would cease to foster general knowledge in individuals and increasingly encourage individual ability in tightly specified and increasingly complex skills. Each worker would become increasingly proficient at one element of the work of a larger project or process. We are currently at a similar point of

decision with digital curation training. It is in the context of these debates that DigCurV operated.

1.2 Demand from Cultural Heritage Sector

The EC has encouraged the growth of digital information professions with the 2005 launch of its i2010 strategy and a subsequent Digital Agenda initiative, launched in 2010.⁴

This investment is justified by the importance of the cultural heritage sector in the European economy. Specifically, in addition to the thousands of universities, libraries and archives across Europe, there are also more than 19,000 museums and art galleries, which employ around 100,000 staff⁵. Traditionally, museums and gallery staff have been trained in physical object care by well-established professional and vocational training courses, but as digital technologies infiltrate every aspect of society, digital objects are increasingly making their way into the collections held by memory institutions.

In 2004, the Digital Preservation Coalition and JISC established the need for digital preservation skills training in multiple sectors in the UK JISC and DPC Training Needs Analysis⁶, and DigitalPreservationEurope research has also echoed the need for these skills to be regularly refreshed by professionals as digital curation practice develops and evolves⁷. In 2009, the New York Times recognised the growing demand for digital archivist skills in the USA⁸. In 2010, Gartner Research identified four new roles needed by IT departments to remain effective⁹ – one of these was ‘digital archivist’, and it was estimated that fifteen percent of businesses would employ in this role by 2012. And yet, at the 2011 JISC ICE Forum in the UK¹⁰, fewer than half a dozen UK institutions were listed as providing digital curation training as part of their profession library and archive courses. The Digital Preservation Coalition is running again in December 2013 its popular course on ‘Getting started in digital preservation¹¹’ and indication of the need and requirement from the sector for basic, easy access training.

The existence of such courses evidences that it is not enough to trust new recruitment into the cultural heritage sector to face the challenges of digital curation. Research conducted by DigCurV confirms that at least in the experience of our respondents, investment is not always channelled towards creating new staff to take on the emerging digital curation duties increasingly required by heritage institutions. There is a need for existing staff to adapt to the emerging digital cultural sector.

2. Methodology

Our paper will describe the research activities of two European wide surveys, focus groups and skills analysis that developed an evaluation framework which was a basis for the curriculum framework. We will focus on the final version of this framework and discuss the research findings that underpin the three lenses, including concept map and model.

The DigCurV Curriculum Framework was developed to compare, describe, and inform the development of training offerings. Useable directly by the individual learner, it can also assist with direction-setting for CPD. The Framework draws on knowledge, expertise and research developed within DigCurV and related initiatives in order to synthesise a matrix of core digital curation skills and competences and, where appropriate, pathways of skills progression between one type of professional role and another.

The DigCurV Curriculum Framework was iteratively developed through extensive testing and evaluation: in the first place, through a series of workshops organised in several locations across Europe; then through a panel of experts in vocational training at a multi-stakeholder workshop, supplemented by targeted interviews and small focus groups with individual professionals. The content of the Framework has been elicited from the professions it describes in accord with its ambition to be genuinely useful to professional practice. In this way, the Framework in its current form provides a robust description of the digital curation professions at the time of publication.

To this end, the Framework comprises three interrelated parts:

- a core Curriculum Framework model, which provides in a cogent, relevant and approachable manner the constituents and interactions of different layers involved in digital curation training;
- three ‘lenses’, or views, one each for three broad types of professional role: Practitioner; Manager and Executive;
- a technical specification which outlines the groundwork for the Framework, defines the Framework’s terminology and identifies the interactions between the Framework and lenses¹².

In the DigCurV context, the cultural heritage sector is understood to comprise museums, libraries, galleries, archives plus relevant departments of HEIs – critical collaborators in digital humanities. The types of training relevant to the project were vocational training for those aiming to enter the profession (including Master’s-level qualification) or those already in post (such as in-house skills training, CPD).

The Curriculum Framework has the capacity to be useful to various audiences, including those working in digital humanities and cultural heritage professions who would like to increase their expertise in digital curation.

Current Use Cases

Various institutions in the higher education sector have found the Framework useful to date. Amongst them, University of London Computer Centre (ULCC), providers of the vocational Digital Preservation Training Programme (DPTP) mapped their curriculum to the Framework, helping to review and reflect on programme content and delivery style. The Department of Information Studies at University College London found the Framework helpful as a tool for skills auditing with those Master’s students who had undertaken an option in Digital Curation. The Framework has also been useful to the University of Aberystwyth in devising its MSc Digital Curation programme. The Professor of Library Science at Purdue University Libraries reported that the Framework has been helpful in understanding the impact of various aspects of the curriculum and the importance of understanding the needs of various professional audiences¹³. This case work in HEIs, rich in digital humanities activities, embeds the DigCurV framework firmly within the DH context.

References

1. ec.europa.eu/education/lifelong-learning-programme/ldv_en.htm
2. www.digcur-education.org/eng/About/Founding-Partners
3. **Glasgow Digital Humanities Network:** www.digital-humanities.glasgow.ac.uk/about
4. **European Commission’s Europe’s Information Society webpage** available at: ec.europa.eu/information_society/eeurope/i2010/index_en.htm.
5. **European Group on Museum Statistics estimate.** Data is available at www.egmus.eu/index.php?id=88&no_cache=1.
6. **JISC and DPC Training Needs Analysis:** www.jisc.ac.uk/media/documents/programmes/preservation/trainingneedsfinalreport.pdf.
7. **Harvey, R.** (2007). *Professional Development in Digital Preservation: a life-long requirement*, DPE briefing paper
8. **De Aenlle, C.** *Digital Archivists: Now in Demand*, New York Times, 7th February 2009. Available at www.nytimes.com/2009/02/08/jobs/08starts.html.
9. **Gartner Identifies Four Information Management Roles IT Departments Need to Remain Effective**, press release available at www.gartner.com/it/page.jsp?id=1282513
10. www.jisc.ac.uk/whatwedo/programmes/preservation/iceforum
11. **European Commission’s Europe’s Information Society webpage** available at: ec.europa.eu/information_society/eeurope/i2010/index_en.htm.

12. Molloy, Gow et al (2013) www.digcur-education.org/eng/Resources/D4.1-Initial-curriculum-for-digital-curators

13. The three latter reactions are available in full in Molloy et al. (2013). Towards an initial curriculum for digital curators. DigCurV project deliverable D4.1, available at www.digcur-education.org/eng/content/download/10836/166027/file/D4.1_Curriculum.pdf

Digital approaches to understanding the geographies in literary and historical texts

Gregory, Ian

Lancaster University, United Kingdom

Donaldson, Chris

Lancaster University, United Kingdom

Murrieta-Flores, Patricia

Lancaster University, United Kingdom

Rupp, C.J.

Lancaster University, United Kingdom

Baron, Alistair

Lancaster University, United Kingdom

Hardie, Andrew

Lancaster University, United Kingdom

Rayson, Paul

Lancaster University, United Kingdom

This paper reports on recent research that explores how geographical information systems (GIS) and related technologies can be used to understand texts, drawing on both literary and historical examples. In *Graphs, Maps, Trees*, F. Moretti identifies mapping as one tool that facilitates distant reading. Other researchers have subsequently demonstrated that GIS can be used to implement this. The research presented here illustrates that the potential for GIS and related technologies in the humanities goes beyond both mapping and distant reading. Specifically, we identify three general ways that geographical technologies can enrich our understanding of texts: first, distant reading using Geographical Text Analysis (GTA), a combination of techniques from GIS-based spatial analysis and from corpus linguistics; second, enhanced close reading based on using place names or maps as query tools; and third, geographical analyses of the texts using techniques such as network analysis and route reconstruction. The aim of these three approaches is to go beyond simply producing visualisations, and instead to allow us to improve our understanding of the text with an emphasis on its geographies.

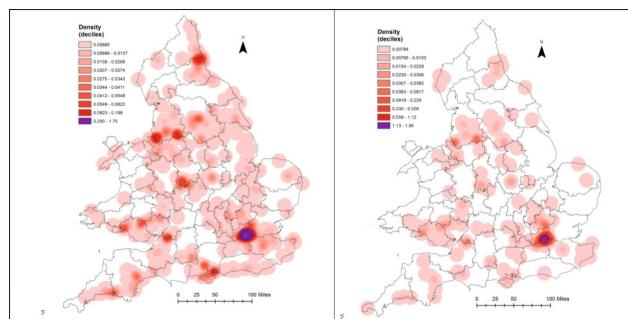


Fig. 1: Cholera in the Registrar General's reports showing (a) locations of cholera instances and (b) instances of terms associated with the water supply.

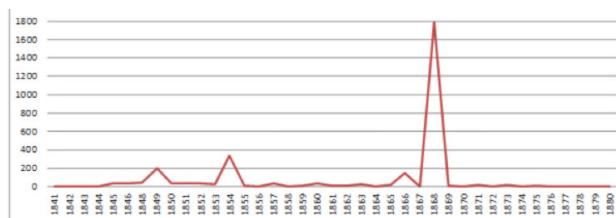


Fig. 2: Time series of cholera instances from the Registrar General's reports. The 1868 spike in instances was not matched by a corresponding rise in deaths.

Distant reading through GTA effectively allows us to ask two basic questions: what places is the corpus talking about? and what places does the corpus relate to a particular theme? This involves more than simple mapping. First, place-names have to be identified using automated techniques. Once this has been done spatial analysis and corpus linguistics techniques allow the geographies within the text to be investigated either in an exploratory way that asks 'where is the corpus talking about?' and 'what is it saying about these places?', or in a more thematic way that asks 'where the corpus is talking about in relation to my theme?' and 'what else is being said about these places?'. As all of the place-names are georeferenced, we are also able to integrate them with other sources that are also georeferenced. To illustrate the potential of this we use the Registrar General's Reports, which document mortality and disease in England and Wales from 1851 to 1911. Using GTA to explore the Registrar General's reporting of cholera showed a number of interesting things: first, that he was particularly interested in cholera in London (figure 1a); second, that the discourse on cholera in London was strongly associated with potential causes, particularly the water supply (figure 1b), whereas in other parts of the country he tended simply to acknowledge that cholera was occurring, increasing or declining; third, that the emphasis on London could not be justified either by the numbers of deaths from cholera in London or London's death rate from the disease; and fourth, that whereas early spikes in instances of cholera in the Reports correspond with known cholera epidemics, the last large spike in 1868 (figure 2), was largely associated with the fear of an epidemic spreading to Britain. Given that there were relatively few cholera-related deaths reported in 1868, we have concluded that the improved understanding of the disease had led to improved measures to prevent it.

These types of distant reading techniques can also be applied to literary texts. Using a corpus of writing about the English Lake District we can show that whereas William Wordsworth was associated with a few central parts of the region in the Romantic period, Victorian readers associated him with sites throughout the Lakes. Using digital images from Flickr, furthermore, we can show that this trend has been reversed in the 20th century.

These top-down, automated techniques are valuable because they allow us to understand large corpora quickly, but they do so at the expense of losing much of the subtlety and nuance that close reading can offer. It is frequently argued that one of the key advantages of digital texts is that they can be read in a non-linear manner. A weakness of this is that it is not always clear how to structure non-linear reading. Place offers one way in. The decline in mortality, particularly among infants (aged under one), started in the nineteenth century but is poorly understood. Much of the research that has been done focusses on the problems and solutions of large urban centres such as London. This is despite the fact that quantitative evidence shows that some rural areas started to decline far earlier than urban centres and at much faster rates. Despite this, there could be major variations between nearby rural areas with apparently similar quantitative characteristics. To explore this further, three neighbouring districts in rural Suffolk - Sudbury, Samford and Risbridge - were analysed. Sudbury and Samford both had relatively high infant mortality rates in the 1850s, the earliest decade for which data are available, but showed rapid improvements thereafter. Risbridge, by contrast, started with low rates, but only showed slight improvements through the rest of the century. In order to explain these variations

we first had to identify all place-names within these districts. This was done using a GIS of the boundaries and a gazetteer. These were then used to query the British Library's Nineteenth Century Newspaper corpus, which contains text from over two million newspaper pages. Additional search terms thought to be relevant to infant mortality decline were used to narrow the searches and this list was refined as the research progressed. Based on the articles found through these queries, we have concluded that the system of local government in Risbridge was far less effective than the systems in the other two districts. Despite many calls to improve drainage, housing and a range of other features that have well established links to infant deaths, little action was taken by Risbridge's authorities. This can clearly be contrasted to the situation in the other two districts, where the local authorities took extensive action. Although this is not a definite causal link, it does provide strong evidence that local government played an important role in reducing rural mortality rates, something that has previously only been identified at the national level or for major urban centres.

Again, similar techniques can be used in literature. We demonstrate this using map-based queries rather than place-names. A system was created that uses a Google Map to show every place mentioned in our corpus of Lake District writing as a point. Each point was linked to web-pages that include the full text. Clicking on a point on the map, presents the reader with a keyword-in-context list (or place-name-in-context) list of all of references to that place and hyperlinks can then be used to follow from these to the appropriate location in the full text. This allows the reader to query not only what is being said about a particular place, but also about nearby places.

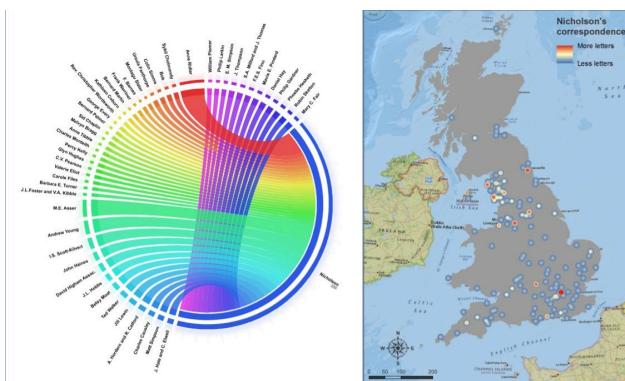


Fig. 3: Network analysis of Norman Nicholson's work. The diagram on the right shows the number of letters sent by Nicholson with thicker lines indicating more letters. The map on the right shows where recipients living in Britain lived.

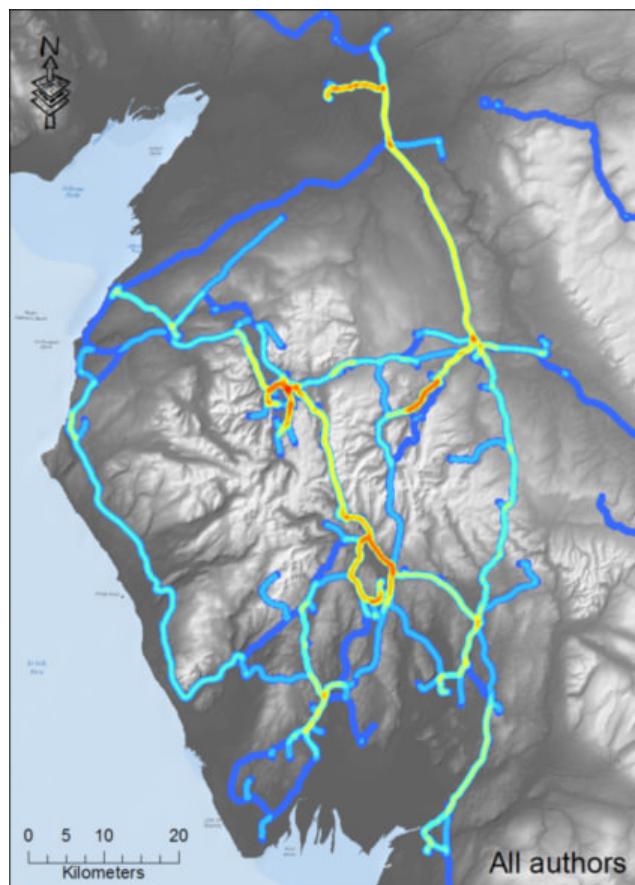


Fig. 4: Cost surface analysis showing the combined estimated routes of Arthur Young (1770), Thomas Gray (1775) and Thomas Pennant (1771 and 1776). Reds and yellows indicate frequented routes.

Finally, geographical technologies can also be used to enhance texts in a number of ways. One way, shown in figure 3, is network analysis which can be used to explore, for example, networks of correspondence. We have used this to explore the correspondence networks of Lake District writers such as Norman Nicholson where a combination of diagrams to show who he was corresponding with and in what volumes, and maps to show where they lived was used. A different approach allows us to move beyond seeing places within texts as isolated points and instead to explore them as parts of journeys. This was done using a number of accounts of journeys through the Lake District. First, the texts were close-read to identify the order in which place-names mentioned were visited. These were mapped as points which were then used as the input into a technique called cost-surface analysis which estimates the most likely route between points. This has been shown to be particularly effective in upland areas such as the Lake District (figure 4). This allows us to estimate and map the routes the writers are likely to have taken, and to explore the geographies of silence concerning the places which writers are likely to have visited but have not mentioned.

In conclusion, the use of geographical technologies in understanding texts is potentially multi-faceted and goes far beyond producing maps. It is instead a useful tool for understanding and enhancing texts to produce the abstract summaries required for distant reading, to select parts of the text that require close reading, and to allow new forms of analyse to help understand the geographies within texts.

Acknowledgements

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). We are also grateful to Sarah Hastings for her work on Suffolk while on an internship.

from Mt Holyoke College hosted in the Department of History, Lancaster University.

References

- F. Moretti**, *Graphs, Maps, Trees*(London, 2005)
- I.N. Gregory and A. Hardie A.** "Visual GISing: Bringing together corpus linguistics and Geographical Information Systems" *Literary and Linguistic Computing*, 26 (2011), 297-314
- C. Grover, R. Tobin, K. Byrne, M. Woolard, J. Reid, S. Dunn, and J. Ball** "Use of the Edinburgh Geoparser for georeferencing digitized historical collections" *Philosophical Transactions of the Royal Society A*, 368 (2008), 3875-3889.
- These are taken from the Histpop collection, see <http://www.histpop.org>.
- D.J. Cohen and R. Rosenzweig**, *Digital History* (Philadelphia, 2006)
- I.N. Gregory** "Different places, different stories: Infant mortality decline in England & Wales, 1851-1911" *Annals of the Association of American Geographers*, 98 (2008), 773-794. <http://www.bl.uk/reshelp/findhelpstype/news/newsdigproj/database/>
- S. Szreter** "The importance of social intervention in Britain's mortality decline c. 1850-1914: a re-interpretation of the role of public health" *Social History of Medicine*, 1 (1988), 1-37

Topotime: Representing historical temporality

Grossner, Karl

Stanford University, United States of America

Meeks, Elijah

Stanford University, United States of America

Topotime: Representing historical temporality

Historical analysis within any discipline depends in part upon establishing chronologies (Sewell 2005), but historical data are problematic. The spreadsheets and database systems used for representing and computing over digital chronologies do not handle vague or otherwise uncertain data well. How does one encode "for 6 months before the war," "around 1832," or "during harvest seasons in her youth?" And when dates for events lasting days or months are given only in years, how can we calculate contemporaneity? What if our data include both precise dates and vague date ranges with varying granularity?

To date, historical researchers and digital humanities application developers have managed temporal uncertainty in ad hoc fashion, normally with one or two date fields using the ISO-8601 standard (e.g. YYYY-MM-DD) for the Gregorian calendar—most would say with less than optimal results. Meanwhile, researchers in computer science, geographical information science, and other fields have done considerable work on some challenges in temporal representation, including uncertainty and qualitative temporal reasoning. (cf. Kauppinen et al 2010; Holmen & Ore 2009; Crescioli, D'Andrea Niccolucci 2000; Plewe 2002). Some of those results, which often demonstrated only in small exemplars, can be brought to bear on humanists' requirements. However, we must make explicit our desiderata for temporal representation and computation in order to make headway towards fulfilling them.

In light of this, we have initiated the *Topotime project*, with these initial goals: 1) a specification for computable digital representations of the kinds of temporal entities typically found in historical texts, and some relationships between them; 2) one or more graphical timeline layout programs to parse and render data written in that form; 3) software tools to facilitate the encoding process, and to transform data in two stages—converting spreadsheet exports to the flexible and human-readable data format, JSON-LD, then parsing and

transforming those JSON data files into "temporal geometries" for calculations of distance, similarity, and topological relations. In this paper, we briefly outline the draft Topotime specification as it stands, and the software development progress we have made so far.

The goals as set out above are admittedly ambitious. Temporal entities found in historical texts and records include a wide range of scales and imprecision, and refer to many calendars and modes of temporal reasoning. We have begun what will be a long-term iterative process of enumerating examples and fine-tuning a data model to handle them, written in JSON.

Two perspectives

We are approaching this work from two directions in parallel to meet requirements for both drawing timelines and for calculating temporal relations. These are not mutually exclusive and parsers for both cases have considerable overlap of functionality and comparable complexity; generalities in formal representation that are useful in both cases are emerging. For example, both parsers convert various date expressions to Julian dates for calculations. There are also distinct differences, for example between the data objects best suited for efficiently drawing time bands, dots, and arrows on a timeline, and the temporal geometries referred to earlier.

The basic elements of Topotime

A Topotime data file describes a *PeriodCollection*. Each *Period* is of a class (either *Event*, *HistoricalPeriod*, or *Lifespan*) and has temporal extents described by one or more typed timespans (*tSpan*). *PeriodCollections* have *Projection* definitions which include atom (granularity, such as day or year), *origin* (day zero on the reference calendar, in Gregorian date terms), and *scale* (used for timeline rendering). Periods must have a unique id, a source attribution, and a label for graphical display. They can also have any number of optional properties (attributes), although Topotime software does not handle these directly. A *PeriodCollection* can also include a set of asserted *relations*, both between periods and between periods and places. These are distinguished from those purely temporal relations between Period timespans, which can be calculated and may be incidental.

Timespans



Fig. 1: Timespan with fuzzy interval bounds, as a probability function. This event likely ended by D (~0.7) and certainly by G.

When describing the "when" of an occurrence we ordinarily mean that it took place either *throughout* some timespan, or for *some time during* it. Someone born in 1723 was not born for the entire year! In Topotime, a *tSpan* describes temporal extents and *throughout* is the default; *some time during* is noted by adding a ("during": True) statement, and a duration, e.g. ("d": "1d") for a duration of a birth day. In both cases (throughout and during), date ranges describe bounds with a required start ("s") and optional latest-start ("ls"), earliest-end ("ee"), and end ("e"). This conforms to a pattern commonly seen in graphical representations, from Joseph Priestley's 18th century timelines to the popular Simile Timeline software[3], and the recent formalizations cited earlier. In this way, timespans can be represented as having either fixed or "fuzzy" bounds[4].

The result is a “temporal geometry” such as pictured in Figure 1. The shapes of the curves between s-ls and ee-e can be articulated more completely by adding “sls” and/or “eee” arrays, as shown, reflecting an author’s understanding of the probabilities over the course of those sub-spans.

Time values for s, ls, ee, and e can be either a day, a month, or a year, and can be qualified by operators for “before” (<), “after” (>), and “about” (~). They can also be pointers to other Period timespans or parts thereof. For example, “>38.s” refers to “after the start of Period 38 in this collection.” Omission of a referent part (e.g. >38 or <38) is taken to mean either its end (“e”) in the first case or its “s” in the latter.

Topotime recognizes not only *date ranges* with fixed or fuzzy bounds, but *durations*, *cyclical timespans* (regularly recurring ranges), and *multi-spans* (arbitrary discontinuous spans) as well. Examples of notation for these are listed in Table 1.

Table 1 - Timespan notation in Topotime, partial listing

fixed range (throughout)	{"s": "1901-04-01", "e": "1963-01-12"} <i>A lifespan: born April 1, 1901; died Jan 12, 1963</i>
fuzzy range (throughout)	{"s": "1923-03-21", "ls": "1923-06-20", "e": "1930-10-01", "ee": "1930-12-31"} <i>Employed from spring of 1923 to late 1930.</i>
fixed range (during)	{"s": "1934", "during": True, "d": "4m"} <i>Traveled in Spain for 4 months in 1934</i>
fuzzy range (during)	{"s": "1923-03-21", "ls": "1923-06-20", "e": "1930-10-01", "ee": "1930-12-31", "during": True, "d": "~6m"} <i>Hospital stay for about 6 months during studies</i>
cyclical	{"s": "1951-05-01", "e": "1999-05-01", "cduration": "18m", "cstep": "4y"} <i>US Presidential campaign seasons in late 20th century</i>
multi-part	[{"s": "1901-01-01", "ls": "1901-02-02", "ee": "1919-01-01", "e": "1920-05-05"}, {"s": "1931-01-01", "ls": "1935-02-02", "ee": "1961-01-01", "e": ">12"}] <i>intermittently as specified, until after Period #12</i>
duration	{"s": ">1", "duration": "2m"} <i>tSpan for Period beginning after Period 1, lasting 2 months</i>

Period relations

Meronomic (parts)

Purely temporal relationships between Periods (overlap, adjacency, containment) can be calculated from *tSpans* geometrically. But there are more relationships we routinely assert and represent, for example *part-of*. We might say, “these 18 events occurring at these times, or in this order, were part-of that larger event”—e.g. a lifespan, war, or political campaign. Another scholar’s chronology for the same composite event might include an entirely different set of sub-events. We may wish to model The Bronze Age as an historical period having spatial-temporal parts such as “Late Bronze Age Southern Levant” and “Bronze Age – Malta.” A Topotime *relation* consists

of a subject, predicate and object (at minimum) in the following form:

{"subj": 23, "pred": "has-part", "obj": 14}

Among other things, Topotime part-of relations enable rendering sub-events within parent containers on timelines.

Time and place

Events and other occurrences are wholly bound to places. Parenthetically, we would argue that places may be best characterized by what has occurred in them. Certainly historical periods are often defined geographically, or are relevant only in particular regions. Some are equally geographic and temporal constructs, e.g. “Pre-dynastic Egypt” (4500-2950 BC), or “The Neolithic Levant.” “Song Dynasty in the Third Imperial Period” is relevant in China and neighboring places, but not elsewhere.

Furthermore, simple and complex events all have spatial extension we often want to display on a map alongside a timeline. Period locations in Topotime can be specified with a single spatial location expressed in a standardized format (GeoJSON or WKT), and with an optional name in this form:

{"subj": 23, "pred": "has-location", "obj": {"name": "Venice", "geom": "POINT (45.4375, 12.3358)", "geomType": "WKT"}}

Standards for specifying Places in data objects like these for gazetteers are now emerging, thanks to the coordinating efforts of projects like Pelagios and national historical GIS projects Great Britain Historical GIS and the China Historical GIS.

Participation

Periods having “class”: “Lifespan” can be asserted to participate-in other kinds of periods, such as Event and HistoricalPeriod.

Looking ahead

Topotime is an open source software development project. We know that many further challenges exist for representing not simply time, but temporality in digital humanities works. Our goal has been to help initiate what will hopefully be an ongoing collaborative process with some concrete steps and functioning software. We are hopeful this work will contribute to the development of interoperable gazetteers of place and period, temporal extensions of the popular GeoJSON format, and improved capabilities for timeline visualizations.

References

- Crescioli M., D’Andrea A., Niccolucci F. (2000). *A G/S-based analysis of the Etruscan cemetery of Pontecagnano using fuzzy logic*, in G.R. Lock (ed.), Beyond the Map: Archaeology and Spatial Technologies, Amsterdam, IOS Press, 157-179.
- Holmen, J., and Ore, C. (2009). *Deducing event chronology in a cultural heritage documentation system*. In CAA 2009 Proceedings retrieved 29 Jul 2013 from www.edd.uio.no/artiklar/arkeologi/holmen_ore_caa2009.pdf
- Kauppinen, T., Mantegari, G., Paakkarinen, P., Kuittinen, H., Hyvonen, E., Bandini, S. (2010). *Determining relevance of imprecise temporal intervals for cultural heritage information retrieval*. International Journal of Human-Computer Studies 68 (2010) 549–560
- Plewe, B. (2002). *The Nature of Uncertainty in Historical Geographic Information*. Transactions in GIS, 6(4): 431-456.
- Sewell, W. (2005). *Logics of History*. Chicago: University of Chicago Press
dh.stanford.edu/topotime ; github.com/ComputingPlace/Topotime
JavaScript Object Notation for Linked Data (json-ld.org).
www.simile-widgets.org/timeline
Note that the term “fuzzy” means indeterminate and probabilistic here; this does not correspond with its meaning in fuzzy set theory, as percent membership in a set.

Does colour mean color?: Disambiguating word sense and ideology in British and American orthographic variants

Grue, Dustin

dustin.grue@gmail.com

University of British Columbia

The orthography/identity hypothesis proposes that a speaker's motivation for selecting between available orthographic variants in a language (e.g. between British *colour* and American *color* in English) is to some extent informed by the speaker's desire to express a certain identity^{1 2 3}. In the Canadian context, where a mixture of American/British variants are used – often with non-categorical preference^{4 5} – Heffernan et al.⁶ developed a method to qualify the orthography/identity connection in terms of ideology and show that during periods of increased “anti-Americanism,” specifically during unpopular American-led wars, American variant use declines relative to the British. Heffernan et al.’s data cover the years 1921 to 2004, and are derived from the student newspaper *The Gateway* at the University of Alberta in Alberta, Canada. Their method involved locating expressions of national sentiment for each year of the data, rating “anti-American sentiment” on a 7-point Likert scale (255 ratings over 85 years performed by each author) and correlating this with the relative frequencies of 15 orthographic variables (Table 2, though *color* / *colour* is my addition): the negative correlation obtained was quite high, with Pearson-r -0.715, $p = 0.001$.

However, follow-up work by the present author, using data in the same timeframe from the archive of the University of British Columbia’s student newspaper *The Ubyssey* (~50 million words) in the neighbouring province of British Columbia, failed to find similar short-term diachronic changes in variant use correlated with periods of increased “anti-American sentiment”⁷. Following Heffernan et al.’s method, an insignificant correlation was obtained: Pearson-r -0.434, $p = 0.064$. Historical relative orthographies do differ between Canada’s provinces^{8 9 10} but, assuming the strong connection between orthography and identity, no clear explanation remains for the lack of correlation in other Canadian data. Without dispensing with the orthography/identity hypothesis, I hypothesize that proximal linguistic contexts are also motivating factors in variant selection, and propose to integrate a more context sensitive model into this top-down, language-external theory of linguistic identity performance.

To test for contextual differences, I treated the problem as one of word sense disambiguation, where the goal is to distinguish lexemes using a set of features and a computational language model—a technique most often used to distinguish between ambiguous meanings of homonyms (such as *judge*, *bank*, *bow*, etc.)¹¹. Features used were a window of words surrounding each variant (8 words either side of the target was found optimal, excluding other instances of variables if present) and the model was Naïve Bayes. If orthographic variants can be discriminated based on surrounding context, we can assume that those words are in some way unique—with the interesting implication, in the extreme case, that spelling variants might not just diverge orthographically but semantically, as well¹². Maybe they mean different things. My experiments attempted to disambiguate variants in each variable from one another using unsupervised and supervised classification, in both cases using the Naïve Bayes form.

Though Naïve Bayes makes the linguistically improbable assumption of feature independence, it has been noted for its precision in classification problems in spite of this simplification (i.e. its ‘naïveté’)¹³. For unsupervised classification I used my own Python implementation of a Naïve Bayes classifier where the parameter estimates are learned through Estimation Maximization (EM), as described in e.g. Manning and Schütze¹⁴. As Pedersen¹⁵ observes, testing the results of unsupervised classification is complicated by the fact that the algorithm

does not assign labels to inputs, instead clustering them, but accuracy can be represented as the proportion of the dominant variant in each cluster. My classifier outputs two ‘sense groups’, which would ideally correspond to the American or Canadian variant. After performing 10 trials and averaging the results, I found that only three variables out of 16 (Heffernan et al. exclude *color* / *colour*—I include it) produced significant results, in that their prediction accuracies departed from – or improved upon – their ‘lower bound’ accuracies, where the lower bound is the relative frequency of each variant and therefore the accuracy one would achieve simply by assigning each variant to a category based on its occurrence. For brevity, only these three are represented in Table 1.

American variant	accuracy	lower bound	Canadian variant	accuracy	lower bound
color	55.5%	54%	colour	65.6%	46%
gray	23.2%	17%	grey	86.2%	83%
jewelry	51.3%	49%	jewellery	55.4%	51%

Unsupervised classification for *colour* improves accuracy by 19.4% over its lower bound, but increases for other variables are marginal and – like *colour* – generally apply to one variant only.

For supervised classification I used the Naïve Bayes classifier in the Python library Natural Language Tool Kit (NLTK)¹⁶, a similar method to Mahowald’s¹⁷ recent study in which *y-* and *th-* pronouns were disambiguated in a corpus of Shakespeare’s plays based on context. Whereas unsupervised classification performed poorly, supervised classification obtained surprising accuracy for multiple variables after 10 validation trials. These results are summarized in Table 2, ranked by accuracy, with an asterisk denoting significance at the $p = 0.001$ level. In this experiment, the lower bound for each variable is 50% because a random subset of the tokens was evaluated and counts were set equal for each variant during testing.

variable	accuracy	total count
jewelry / jewellery	81.6%*	564
gray / grey	79.3%*	3112
color / colour	74.5%*	5312
program / programme	70.1%*	5704
honor / honour	62.0%*	2538
enrollment / enrolment	61.2%*	2086
humor / humour	61.1%*	2862
neighbor / neighbour	60.7%*	494
defense / defence	58.6%*	5640
judgment / judgement	56.8%	928
offense / offence	56.3%*	1568
centered / centred	55.7%	488
marvelous / marvellous	55.6%	316
fulfill / fulfil	54.7%	312
labeled / labelled	54.4%	270
kilometers / kilometres	40.0%	72

It would seem that we are able to predict, sometimes with high accuracy, whether certain variables will realize their

American or British variant based on context. But why? If orthographic variations are simply different graphemes of the same lexeme, decided rather capriciously by an American lexicographer in the nineteenth century¹⁸, why should this be possible?

The Naïve Bayes module in NLTK provides output for identifying features most useful in making its decisions, and can help answer this question. For *gray / grey*, the case is clear, since the terms most likely to indicate British *grey* are *Pt.* and *Point* (Point Grey is the name of the land on which the University of British Columbia lies), and terms indicating American *gray* are proper nouns like *Bob*, *John*, and *Stuart* (*Gray* is a common surname). A revealing result, but only so far as it reveals a highly restrictive context in non-compositional forms, and might suggest this variable be excluded from further testing. Contexts for *color / colour* are more interesting, however, and fall into two large subjective categories: ‘cultural’ and ‘technological’ (Table 3).

variant	category	informative features
colour	cultural	diversity, women, people, racism, queer
color	cultural	people
colour	technological	connected, jet, print, modem, monitor
color	technological	cartoons, TV

As a collocation analysis reveals, in the ‘cultural’ category phrases such as *women of colour*, *people of colour*, and *queers of colour* occur often with *colour* – 225 total instances, its most frequent collocate – but hardly ever with *color* (11 instances). In the ‘technological’ category, computer terms appear with *colour* and entertainment terms with *color*, where these terms are often found in advertisements and the site of these interactions tend to be local for *colour* and global for *color* (a local transaction for a *colour monitor*, at least prior to the expansion of current global markets, but the international consumption of *color television*). However, these phrases are easily recognizable as historically specific (to around post-1980). Indeed, the unsupervised classification of *colour* backs up this historical selectivity: significantly more of the items grouped at 65.6% accuracy are from this decade—context and history are intertwined, of course, and it exceeds my scope to disambiguate these here. But the more a-historical distribution of terms predicting *jewelry / jewellery* suggests historical clustering is not inevitably the rule: local activities like *piercing* and *repairs*, and localities denoted by *West* and *Point* (i.e. the location of a shop in *West Point Grey*) predict British *jewellery*, but generic sales terms *accessories*, *fine*, *place*, and *giftware* predict American *jewelry*.

In sum, advertisements, or, more generically, ‘solicitations’, are the dominant vehicle of these variants and prefer the British when the activity is local (both economically and socially—these will be further described) and the American generally. I will also further discuss how accounting for genre affects classification accuracy. Overall, British variants are more uniquely contextualized, and therefore more easily discriminated, than American.

Qualitative sociolinguistic approaches like Heffernan et al.’s (2010) locate identity as an exterior motivating condition for language, with the necessary assumption that orthography is selected independently of linguistic context. And though this paper finds that this assumption does not hold, the ability to disambiguate orthographic variants based on context is interesting, but not explanatory in its own right. These contexts are also motivated, and computational techniques take us full-circle back to considering ideological – but more interactional – motivations for linguistic context.

References

1. Lipski, J. (1975). *Orthographic variation and linguistic nationalism*. La Monda Lingvo-Problemo 6. 37-48.
2. Schieffelin, B. B., and Doucet, R. C. (1994). *The ‘real’ Haitian Creole: Ideology, metalinguistics and orthographic choice*. American Ethnologist 21. 176–200.
3. Sebba, M. (2000). *Orthography and ideology: Issues in Sranan spelling*. Linguistics 38. 925–948.
4. Chambers, J. K. (2011). ‘Canadian dainty’: The rise and decline of Criticisms in Canada. In Legacies of Colonial English: Studies in Transported Dialects, 224-241. Cambridge: Cambridge UP.
5. Pratt, T. K. (1993). *The hobgoblin of Canadian English spelling*. In S. Clarke (ed.), Focus on Canada, 45–64. Amsterdam: Benjamins.
6. Heffernan, K., Borden, A., Erath, A. C., and Yang, J.-L. (2010). *Preserving Canada’s ‘honour’: Ideology and diachronic change in Canadian spelling variants*. Written Language and Literacy 13(1). 1-23.
7. Grue, D. (forthcoming). *Testing Canada’s ‘honour’: Does orthography index ideology?* Strathy Student Working Papers on Canadian English.
8. Brinton, L. and Fee, M. (2001). *Canadian English*. In John Algeo (ed.), The Cambridge history of the English language, vol. 6, 422-439. Cambridge: Cambridge UP.
9. Ireland, R. J. (1979). *Canadian spelling: An empirical and historical survey of selected words*. Ph.D. dissertation, York University, Toronto.
10. Ireland, R. J. (1980). *Canadian spelling: How much British? How much American?* English Quarterly 12(4). 64-80.
11. Pedersen, T. (2002). *A baseline methodology for word sense disambiguation*. Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics. 126-135.
12. Miller, G. A. and Charles, W. G. (1991). *Contextual correlates of semantic similarity*. Language and Cognitive Processes 6(1). 1-28.
13. Abney, S. (2008). *Semisupervised Learning for Computational Linguistics*. New York: Chapman & Hall.
14. Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, Mass.: MIT Press.
15. Pedersen, T. (1998). *Learning Probabilistic Models of Word Sense Disambiguation*. Ph.D. Dissertation, Southern Methodist University, University Park, Texas.
16. Bird, S., Klein, E., and E. Lopez. (2009). *Natural Language Processing with Python*. Sebastopol, CA: O'Reilly Media.
17. Mahowald, K. (2012). A Naïve Bayes classifier for Shakespeare’s second-person pronoun. Literary and Linguistic Computing 27(1). 17-23.
18. Webster, N. (1846). *A dictionary of the English language*; abridged from the American dictionary. [American Dictionary]. New York: Huntington and Savage.

Navigating the Storm: eMOP, Big DH Projects, and Agile Steering Standards

Grumbach, Elizabeth

egrumbac@tamu.edu

Initiative for Digital Humanities, Media, and Culture at Texas A&M University

Christy, Matthew

mchristy@tamu.edu

Initiative for Digital Humanities, Media, and Culture at Texas A&M University

Mandell, Laura

mandell@tamu.edu

Initiative for Digital Humanities, Media, and Culture at Texas A&M University

Neudecker, Clemens

Clemens.Neudecker@KB.nl
KB National Library of the Netherlands

Auvil, Loretta
lauvil@illinois.edu
Illinois Informatics Institute (I3) at the University of Illinois at Urbana Champaign

Samuelson, Todd
Cushing Memorial Library & Archives at Texas A&M University
todd samuelson@library.tamu.edu

Antonacopoulos, Apostolos
A.Antonacopoulos@primaresearch.org
Pattern Recognition and Image Analysis (PRImA) research Lab at The University of Salford

Introduction

In 2011, the Comite de Sages presented “The New Renaissance” to the European commission, stating that “digitiz[ation] is more than a technical option, it is a moral obligation” to the public. The report stresses that the initiative’s goal is to ensure that we “experience a digital Renaissance instead of entering into a digital dark age.” If the lack of adequate, searchable early-modern digital resources can be correctly referred to as a “digital dark age,” then we are undoubtedly seeing the emergence of a “digital renaissance.”¹ Projects like IMPACT (Improving Access to Text²), eMOP (Early Modern OCR Project), TCP (Text Creation Partnership), and others have emerged in recent years to take up the call to arms issued by the Comite de Sages. However, we need more than the digitization of cultural materials; we need *responsible* digitization alongside a community engaged in the fight for digital visibility of those materials. And, most importantly, large DH projects need effective and responsible management and collaboration standards. The aim is to adapt and adjust to the changing climate, ultimately steering the project safely into the harbor.

Overview

Many OCR (Optical Character Recognition) and cultural preservation projects are underway that need to be able to adapt their project plans to OCR technology and crowd-sourcing breakthroughs as they occur. In Fall 2012, the Initiative for Digital Humanities, Media, and Culture at Texas A&M University received a \$734,000 grant from the Mellon foundation for the Early Modern OCR Project (eMOP)³. eMOP’s objective is to make machine readable, or improve the readability for, 45 million pages of text from two major proprietary databases: Eighteenth Century Collections Online (ECCO) and Early English Books Online (EEBO). Generally, eMOP intends to improve the visibility of early modern texts by making their contents fully searchable. The current paradigm of searching special collections for early modern materials by either metadata alone or “dirty” OCR is inefficient for scholarly research (Mandell, 2013⁴). We intend to publish an open source OCR workflow at grant end in Taverna. This workflow will contain access to an early modern font database, customization guidelines for the Tesseract OCR engine, post-processing and diagnostic algorithms, and crowdsourcing and “scholar-sourcing” (as Brian Geiger has dubbed) correction tools. But the overarching goal of eMOP, a project that blends book history⁵, digital humanities, textual analysis, and machine learning, is ultimately to foster a community of scholars and institutions interested in the digital preservation of, and access to, these texts. To this end, eMOP has assembled an international team of collaborators from multiple disciplines.

eMOP, however, has faced problems in the implementation of our goals and processes. During Year One, the eMOP team and collaborators quickly realized that the grant document excellently outlined milestones and goals, but it did not provide the level of granularity needed to complete each. We have also realized that progress is continually changing in this field, and if big DH projects do not adjust accordingly, they will end

up reinventing the wheel. Active outreach and collaboration with institutions outside the initial grant collaborators proved important. In addition, eMOP is working with proprietary page images and metadata in order to release an open source tool, which has produced its own challenges. In order to succeed in producing a corpus of machine-readable texts and a workflow for future OCR projects, continual outreach and collaboration is needed, yet not always possible due to the restrictions of grant deadlines, funding, and other institutional roadblocks.

Getting Started

This panel considers how big DH projects, with big datasets, big networks of collaborators, and big goals, can adjust and adapt to change. It has long been noted that digital humanities projects lend themselves well to agile⁶ development models⁷, specifically the “the philosophy of ‘releasing early and often’” (Scheinfeldt, 2010⁸). However, these models often break down in the face of multi-institutional and international collaboration, software development, assembling large amounts of data, and what James Smithies and enterprise IT call “transition management,” or planning for “Change” (2011⁹). A digital humanities project, large or small, also “seems to both depend upon collaboration and aim to support it” (Spiro, 2009¹⁰). Each big DH project must find a practical balance in development management *and* collaboration methods.

This panel will bring together the eMOP management team at the IDHMC and collaborators from various disciplines and institutions to discuss the reasons why big DH projects need to plan for adaptation, ways in which projects can achieve this flexibility, and how to swiftly change directions.

If eMOP’s goal reflects of the goal of digital humanities at large, i.e. to foster collaboration among various disciplines and cultivate inter-institutional and international relationships that make possible new kinds of humanities research, then this panel provides a microcosm of that endeavor.

Panel Organization

This panel will consist of a brief 5 minute overview of the goals of, methodologies for, and collaborators in the Early Modern OCR Project, and then each speaker will introduce a major directional change or challenge that eMOP has faced, including the resulting solution in 7 minutes or less. Introductions to challenges may include comparisons to other large dh projects (e.g. IMPACT). Discussion of the resulting solution may include a short software/tool demo. The panel organizers will then pose questions to the roundtable to begin an open conversation, leaving the remaining time for discussion amongst panelists and the audience. Discussion will likely focus on how to change directions, rethink decisions, and reconfigure plans when collaborating with multiple institutions and individuals while facing grant deadlines and milestones.

Questions that Panel Organizers may pose:

- Discuss future models of big DH project management, especially how essential multi-institution and international collaboration can be.
- What can big DH projects learn from the agile vs. traditional software development models?
- Discuss best practices in project management, and how they might be modified in order to take in recent technological innovations or respond to challenges.
- We know that “failure” (Unsworth, 1997¹¹) is important: how can small failures be channeled into big success?
- What kinds of cultural practices need to be taken into account when U.S. projects adopt European models, and vice versa?
- How can transatlantic collaboration best be orchestrated so that projects benefit from collaborators’ advancements, both technological and social?

Participants

- All panelists are committed and eagerly anticipating the discussion of eMOP, large DH projects, and successful and responsible collaboration and development management.
- Apostolos Antonacopoulos is the Director of the Pattern Recognition and Image Analysis (PRImA) research lab at the University of Salford, UK. Dr. Antonacopoulos has been working on issues of pattern recognition, image and document analysis, and historical document digital restoration for many years. In addition to eMOP, he has contributed to the IMPACT and Europeana Newspaper projects, and will discuss how the adoption and customization of software for large cultural preservation projects should be responsive to changing project needs.
 - Loretta Avuil works at the Illinois Informatics Institute (I3) at the University of Illinois at Urbana Champaign. She has worked with a diverse set of application drivers to integrate machine learning and information visualization techniques to solve the needs of research partners. Prior to working for I3, she spent many years at NCSA on machine learning and information visualization projects and several years creating tools for visualizing performance data of parallel computer programs at Rome Laboratory and Oak Ridge National Laboratory. She will be discussing big DH projects from the perspective of these experiences and her work with eMOP.
 - Liz Grumbach is Project Manager for the Advanced Research Consortium (ARC) and IDHMC "alt-ac" Research Staff. She is Co-Project Manager for eMOP (Year Two), and will briefly introduce the project (goals and methodologies). She will also end the panel by comparing the current workflow for the eMOP OCRing process with the proposed OCR workflow contained in the grant, summing up the overall changes that each collaborator's contribution shaped.
 - Laura Mandell is Professor of English and Director of the IDHMC at Texas A&M University. In addition to being the Lead PI for eMOP, Dr. Mandell previously received a Mellon grant (2010) to investigate how effective the open-source OCR engine Gamera could be trained to read early modern fonts. She will introduce the data management challenges eMOP has faced, demonstrating software and tool solutions created by eMOP graduate students and staff.
 - Clemens Neudecker serves as Technical Coordinator in the Research section of the Innovation & Development Department of the KB National Library of the Netherlands. He has been working in numerous large-scale national and international digitization / digital humanities projects since the early 2000's, with a particular focus on OCR (www.impact-project.eu) and scalable workflows (www.scape-project.eu), and will be discussing how this previous knowledge aided the eMOP team.
 - Todd Samuelson is Assistant Professor at Texas A&M University and the Curator of Rare Books & Manuscripts at Cushing Memorial Library & Archives. Dr. Samuelson is the book history consultant for eMOP. He will discuss font history research roadblocks and demonstrate font creation and identification tools created by eMOP collaborators to solve these issues.

Panel Organizers:

- Matthew Christy**, Lead Software Applications Developer for the IDHMC and Co-Project Manager for eMOP (Year Two)
Liz Grumbach, IDHMC "alt-ac" Research Staff and Co-Project Manager for eMOP (Year Two)

References

1. European Commission: *The Comité des Sages. The New Renaissance: Report of the comité des sages on bringing Europe's cultural heritage online*. By **Elizabeth Niggemann, et al.** 10 Jan 2011.
2. **IMPACT**. *Annual Report: Project Periodic Report*. Netherlands: IMPACT, 2011. Improving Access to Text. 9 Dec 2011. www.impact-project.eu/uploads/media/

IMPACT_Annual_report_2011_Publishable_summary_01.pdf .
29 Oct 2013.

3. **Mandell, Laura**. *Mellon Foundation Grant Proposal: "OCR'ing Early Modern Texts."* Grant Proposal. 30 Jun 2012.
4. **Mandell, Laura**. (2013) *Digitizing the Archive: The Necessity of an 'Early Modern' Period*. Journal for Early Modern Cultural Studies 13.2: 83-92.
5. **Heil, Jacob and Todd Samuelson**. (2013) *Book History in the Early Modern OCR Project, or, Bringing Balance to the Force*. Journal for Early Modern Cultural Studies 13.4 (2013): 90-103. Web. 30 Oct 2013.
6. **Beck, Kent, et al.** *Manifesto for Agile Software Development*. Agile Alliance. 30 Oct 2013.
7. **Martin, Robert Cecil** (2003). *Agile Software Development: Principles, Patterns, and Practices*. Saddle River, NJ: Prentice Hall.
8. **Scheinfeldt, Tom**. *Stuff Digital Humanists Like: Defining Digital Humanities by its Values*. Found History. 2 Dec 2010. www.foundhistory.org/2010/12/02/stuff-digital-humanists-like . 30 Oct 2013.
9. **Smithies, James** (2011). *A View from IT*. Digital Humanities Quarterly 5.3. Web. 30 Oct 2013.
10. **Spiro, Lisa**. *Examples of Collaborative Digital Humanities Projects*. Digital Scholarship in the Humanities. 1 Jun 2009. digitalscholarship.wordpress.com/2009/06/01/examples-of-collaborative-digital-humanities-project . 30 Oct 2013.
11. **Unsworth, John**. *Documenting the Reinvention of Text: The Importance of Failure*. The Journal of Electronic Publishing 3.2 (1997). Web. 30 Oct 2013.

Accessing, navigating, and engaging with high-resolution document image collections using Diva.js

Hankinson, Andrew

andrew.hankinson@mail.mcgill.ca
McGill University

Pugin, Laurent

laurent.pugin@rism-ch.org
Swiss RISM/ Fribourg University

Fujinaga, Ichiro

ich@music.mcgill.ca
McGill University

High-resolution page images are providing digital humanities researchers with unprecedented visual access to historically significant works located around the world. As libraries and archives continue digitizing their historical document collections, they are increasing the quality and resolution of their document imaging systems and producing images that, while unprecedented in their clarity and detail, are inconvenient to navigate and manipulate in traditional browser-based environments. Users find themselves waiting for large PDF files to download, or clicking endlessly through thumbnail after thumbnail to find a page image that contains materials of interest to them. These methods of document navigation and viewing have been in place since the infancy of the web browser, and are needlessly awkward given advances in creating asynchronous web applications.

The most common interface paradigm for browsing images online is the 'image gallery'. To illustrate this type of interface we will use the example of the Early English Books Online (EEBO) interface as one with which some readers may be familiar. In the EEBO image viewing interface, users navigate a document as if it were a series of independent images, or image gallery, viewing small thumbnails that, while efficient for downloading to a browser, make it impossible to see the actual content of the page (Figure 1). To examine any single page, the user must click on an image, bringing up a second view of the page optimized for viewing in the browser, but may

not be usable for close examination if the text on the page is too small. Should a user wish to examine any part of a page in particular detail, there may be an option to download a larger, high-resolution image, but the user must wait for this large image to download to their browser, which may take several minutes depending on the speed of the network connection and the size of the image. If the user waits for the full quality image to download but wishes to continue browsing the document on the next page they must traverse back to the smaller thumbnails and start the process again.

View this document as:

Authors: Rowley, William, 1585?-1642?
Title: The birth of Merlin, or, The childe hath found his father **Date:** 1662

View thumbnails: 1-28



Fig. 1: Viewing page thumbnails in the EEBO collection.

An alternative to the image gallery mode of interaction is the use of a browser-based book reader component. Several of these systems are available for managing user interactions with page images. Perhaps the most well-known purpose-built web-based document viewer is BookReader by the Internet Archive¹. Developed as part of the Open Library project, this software presents the user with a book metaphor, inviting them to 'turn' the pages of a book. While this provides a useful alternative to the image gallery mode of viewing, the IA BookReader requires that each page is represented by a complete image file, so zooming in and viewing a page in detail requires the user to wait while the entire image is downloaded—which can be slow and cumbersome depending on the size and resolution of each image. This is also true for PDF-based document image display, where a user is forced to accept a trade-off between viewing low-resolution versions of page images, or waiting for extremely large PDF files to download before they can view any of the pages.

To optimize viewing large, high-resolution documents in a web browser we developed the Diva document image viewer. Diva features several methods for managing user interactions with document page images. Users interact with the full document by scrolling the document, as they might with a PDF file. However, in Diva all page images are composed of smaller tiles. These tiles are of a fixed size (256x256 pixels), and all pages (and their tiles) that are outside of the user's viewport are not downloaded to the browser. This creates an 'instant-on' effect to viewing a document, since the user does not have to wait for the entire document to download, but just a small portion. As a user scrolls, new tiles are downloaded on-demand. To view higher or lower resolution page images, users can 'zoom' between resolutions. Zooming in and out on an image will download just the portion of the page that fits on the users' screen (Figures 2 and 3).

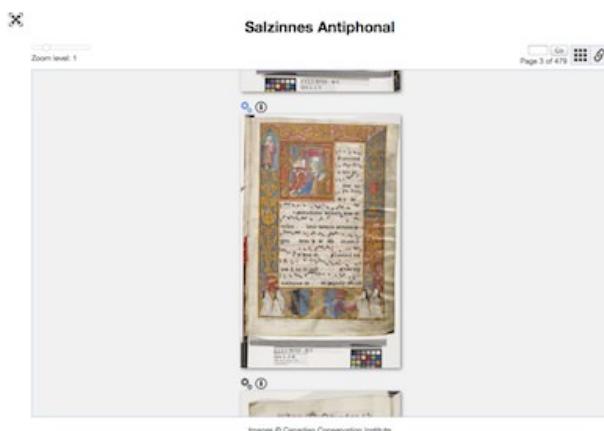


Fig. 2: A zoomed-out view of a manuscript page.

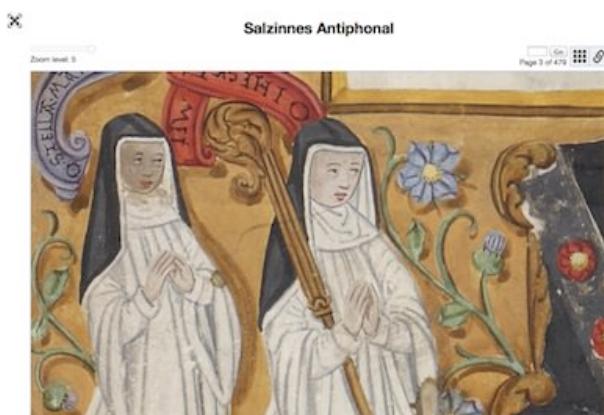


Fig. 3: Zoomed-in detail of the lower-left corner of the page shown in Figure 2.

With the ubiquity of mobile devices it is important to ensure Diva functions on low-memory systems, like the iPad or iPhone. Diva.js uses several methods unique to document image viewers for optimizing memory usage and display. While a document may be several hundred pages long, Diva keeps just three pages in memory at any given point in time, dynamically adding and removing page elements from the browser as the user scrolls. This creates a fast and efficient browsing system for both mobile and desktop devices.

Furthermore we have built a number of image manipulation tools into Diva.js that allow users to engage with a document. Many documents, especially older manuscripts, feature faded inks or text that is written perpendicular to the captured page orientation (e.g., marginalia). Using Diva, users can manipulate brightness, contrast, and page rotation in their browser via an unique set of HTML5-based image manipulation tools, allowing them to enhance faded inks or rotate a page to read margin notes or tables (Figure 4). Other viewers that offer this functionality manipulate the image on the server and then send it back to the client. This is a high-latency operation. With browser-based image manipulation users can see the results of their changes immediately.



Fig. 4: Manuscript page (in Arabic) rotated 90° to view perpendicular text on flyleaf. The controls on the left allow the user to manipulate brightness, contrast, rotation, zoom, and individual RGB colour channels.

We have used Diva.js as the presentation layer of a document image search system. When a user searches for a given word or phrase, the results are presented *in situ* on page images, highlighting the exact location on each page where their result occurs.

All components of Diva are available as free and open-source software, available on GitHub². Diva may be integrated into existing digital library systems. On the server-side, Diva requires the IIP Image Server³, a giga-pixel image server that serves the page image tiles, and a standard web server, such as Apache or NginX. Document images can be encoded as either multi-resolution JPEG2000, or pyramid TIFF files. The JavaScript components of Diva.js will work in any modern web browser. These components manage the asynchronous communication process between the user's browser, the web server, and the IIP Image Server. We have also built in a comprehensive API and plugin system that provides 'hooks' into the page loading and image manipulation systems.

In our presentation we will provide a demonstration of Diva.js and an overview of its background and development history. We will discuss several case studies where we have employed Diva.js for viewing and searching large historical document image collections, where it is used to display page images captured at over 1,200 PPI. Zooming in on images at these resolutions allow users to view individual brush strokes, paper detail and condition, to view details of manuscript illuminations, and several other important document factors that are lost in lower-resolution image displays. Finally we will demonstrate several new features for highlighting and annotating places of interest on page images and integrating page images with the output of optical character recognition software.

References

1. The Open Library. 2013. Internet Archive Bookreader. openlibrary.org/dev/docs/bookreader (accessed 7 March 2014).
2. Distributed Digital Music Archives and Libraries. 2014. Diva.js github.com/DDMAL/Diva.js (accessed 7 March 2014).
3. Pillay, R., and D. Pitzalis. (2008). IIP Image Server. iipimage.sourceforge.net (accessed 7 March 2014).

The Ancient Coins of Thrace: A Numismatic Web Portal

Hanrahan, Elise

walther@bbaw.de
BBAW, Germany

Project Summary

In the field of numismatics there have been many attempts to create overarching reference works for ancient Greek and Roman coins, with different degrees of success. Challenges include the sheer amount of coins to be documented and the difficulty of accessing them, being spread throughout the world. A virtual collection however opens up new possibilities and promises to come closer to this aim through collaboration and linked data.

The Ancient Coins of Thrace concentrates on coins originating from a specific ancient region (Thrace—today's Bulgaria, northern Greece, and European Turkey) digitally collected from around the world. A web portal of this character is quite unique within the numismatic community. The project hopes to provide a method which others can emulate, perhaps leading to a group of online resources which together accomplish that which a printed edition could not. Funding comes from a three year DFG (Deutsche

Forschungsgemeinschaft) grant. The project is located at the BBAW (the Berlin Brandenburg Academy of Sciences and Humanities) and is currently starting its second year. Its main partner is the Bode Museum Coin Cabinet, part of the Berlin State Museums.

The project has two goals. The first is the virtual collection itself: a web portal where ancient Thracian coins can be found. This is accomplished in three ways.

- The BBAW: The academy has its own extensive collection—numbering almost 33,000 plaster casts of about 16,500 Thracian coins. The coin data is being entered into the project data base along with scans of the plaster casts.
- Cooperation with museums and institutions: The project's first and main partner is the Bode Museum Coin Cabinet. They are providing data and photos from about 4,000 Thracian coins. Other museums and institutions have expressed themselves as ready partners, such as the ANS (American Numismatic Society).
- Individual entry: This is one of the project's more special features. Users will be able to register on the web portal and enter their own Thracian coins. This is especially for smaller museums that do not yet have data bases or for private collectors. Smaller museums would then have an online presentation of their Thracian coin collection.

The second goal is research-oriented: it involves the typification of Thracian designs (the pictures on coins) and legends (the texts on coins), together creating the 'type' of a coin. Through standardized design descriptions and legends, the identification of dies² is greatly furthered. All coins found in the portal will thus be linked to these standardized descriptions, and as a result dies can be identified and given a fixed number. This is an essential step for numismatic research.³

What challenges arise from such an undertaking?

Digital standards: One important focus of the project is implementing and promoting numismatic digital standards. To this purpose nomisma IDs⁴ are used for all relevant fields, and lists are being sent to nomisma for new IDs for those which do not yet exist. The nomisma data base standards (NUDS) were also used wherever possible. Such standards are essential for data exchange but will only become truly useful when implemented by other institutions. The Bode Museum Coin Cabinet is participating by entering nomisma IDs for Thracian coins. The ANS, being the main partner of nomisma, have nomisma IDs for their coin data as well. It is hoped through example to encourage the use of these standards in other numismatic data bases.

Import and web interfaces: the original concept for the portal was to import coin data from other institutions into its own data base. This is currently the method in process for the ca. 4,000 coins from the Bode Museum Coin Cabinet. The Bode Museum is however a very involved partner, working actively and closely for a successful import. One could raise the question if perhaps web interfaces—something like windows into different Thracian coin collections—might not be a more practical solution for other institutions. It must not be forgotten, however, that all coin data must still be linked to its standardized description. Such questions are still being discussed.

Individual entry: This rather special feature of the web portal brings its own difficulties. Accessing collections which are not yet in data bases and involving these members of the numismatic community is essential to the project. Holding very high standards for coin data (both in terms of the numismatic research and digital standards) is however equally important. A very strict entry mask could alleviate difficulties but does not offer a final solution. It remains to be seen how much reworking the individually entered coins will require.

Data model: The data model is complex and its conception took time. Yet the even greater challenge was translating numismatic research goals into the language of data modeling. In other words—as an initial step before creating the data model—the challenge of understanding the aims of the research,

the relationships between numismatic concepts, and how the research process itself would enfold.

Conclusion

The Ancient Coins of Thrace strives to create a web portal that offers easy access to Thracian coins, that provides a typification of design and legend, and that furthers numismatic digital standards. It hopes to be of use to museums, private collectors, students and researchers. Because many aspects of the project are relatively new for the field of numismatics (such as a regional based web portal, individual coin entry, and linked data with nomisma IDs), it is certain that these efforts could be improved with time and experience. At the conference we wish to share our progress and developments thus far, receive feedback, and exchange ideas with other projects.

References

1. See Robert Bracey's article 'Online Numismatic Databases – A Review' (2012) for a good overview of what's currently out there: www.academia.edu/2335397/Online_Numismatic_Databases_-_A_Review
2. 'Metal piece engraved with the design used for stamping the coin.' en.wikipedia.org/wiki/Glossary_of_numismatics
3. 'Research needs the die, not the exemplar' (Die Wissenschaft braucht den Stempel, nicht das Exemplar) wrote Mommsen in an expertise „über den wissenschaftlichen Werth ... einer Gesamtpublikation der antiken Münzen“ 1887. Cited by H.-M. Kaenel, Schweizer Münzblätter 2004, here p. 85, see also idem., Theodor Mommsen. Zur wissenschaftlichen Edition antiker Münzen. Gutachten aus dem Jahre 1886, SNR 81, 2002, 7–20, here p. 9.
4. 'Nomisma.org is a collaborative project to provide stable digital representations of numismatic concepts according to the principles of Linked Open Data. These take the form of http URLs that also provide access to reusable information about those concepts, along with links to other resources.' nomisma.org/

The Chimeria Platform: User Empowerment through Expressing Social Group Membership Phenomena

Harrell, D. Fox

fox.harrell@mit.edu

Comparative Media Studies Program, Massachusetts Institute of Technology

Kao, Dominic

dkao@mit.edu

Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology

Lim, Chong-U

cylim

Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology

Lipshin, Jason

lipshin@mit.edu

Comparative Media Studies Program, Massachusetts Institute of Technology

Sutherland, Ainsley

ainsleys@mit.edu

Comparative Media Studies Program, Massachusetts Institute of Technology

1. Introduction

Computational modeling of social categories can be found in a wide range of digital media works. For example, within computer role-playing games (RPGs), racial categorization is often used to style the visual appearance of a player's avatar or trigger different canned reactions when conversing with a non-player character (NPC). In social media, users might join groups based on shared taste or categorize each other into groups such as "colleagues" or "family members" using privacy settings. However, in most such systems, category membership is determined in a top-down fashion. Members are often slotted into single, homogeneous groups, with no possibility for hybrid identities, identities that exist at the margins of groups, or identities that change over time. Taken holistically, such approaches have many limitations. These deficiencies are particularly visible when trying to accurately model the nuance of social category membership in the real world.

Our Chimeria platform (hereafter Chimeria) addresses this deficiency. It creates more nuanced social categorization models in two primary ways: (1) by modeling the underlying structure of many social categorization phenomena with our Chimeria engine; and (2) by enabling users to build their own creative applications about social categorization, using the engine as a backbone. Drawing on theories from sociolinguistics (Polyani, 1989), cognitive science (Lakoff, 1987), and sociology of classification (Bowker and Star, 1999), the underlying engine allows for the movement of individuals within, between, and across social categories. It also allows for members to be more central to a group than others, to assimilate or naturalize in relation to a hegemonic group, and to claim membership in multiple groups. In this paper, we discuss the components of Chimeria and two sample applications built with it.

2. The Chimeria Authoring Platform

Chimeria supports authoring narratives of group membership in any social identity domain through a data-driven approach. Chimeria is divided into three components (Figure 2).

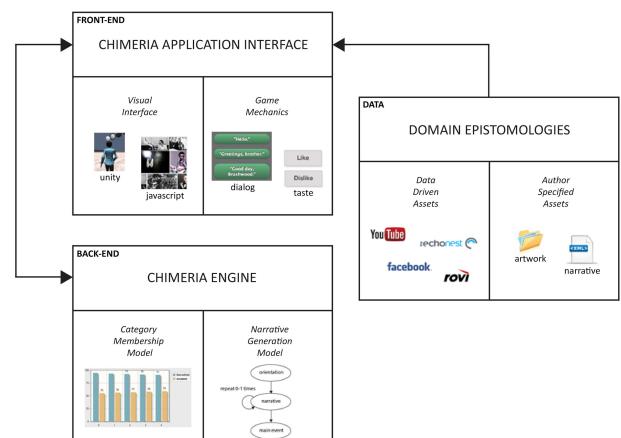


Fig. 1: The Chimeria Platform

1. Chimeria Engine: A mathematical model of users' degrees of membership across multiple categories. It provides the functionality to calculate, modify, and simulate changes to these memberships and serves as the logical processing component of the system. It models users' category memberships as gradient values in relation to the membership values of more central members (Harrell, 2010; Bowker and Star, 1999; Lakoff, 1987). This enables more representational nuance than binary statuses of member/nonmember. Narratives processed by Chimeria are authored using a GUI or in the XML file format with a narrative structure as described in (Harrell et al., 2013).
2. Chimeria Application Interface: A visual interface for user interaction and for experiencing the narratives related to the category membership changes driven by the Chimeria Engine. It provides freedom and flexibility over the aesthetic and visual components of narratives. The interface can take

on multiple forms (e.g., a text-only interface or a 3D virtual environment).¹ The separation between the back-end (the Chimeria Engine) and the front-end (Chimeria Application Interface) provides the flexibility to go through the same narrative trajectory in relation to membership shifts but with varying visual appearance.

3. Chimeria Domain Epistemologies: An “epistemology” is an ontology that describes cultural knowledge and beliefs (Harrell, 2013). In Chimeria, they are the knowledge representations describing the categories being modeled. The data utilized by Chimeria to present these categories to users include both author-contributed (e.g., artworks or narratives) and data-driven (e.g., an API call to YouTube to query for a video) assets.

3. Chimeria Application Domains

To better illustrate the capabilities of the components within our system we describe two very different narratives created using Chimeria: 1) a fictional social networking application which models social categories in the domain of musical preferences (Harrell, 2013); and 2) a computer role-playing game (RPG) scenario which models a conversational narrative between the player and a non-playable character (NPC).

3.1 Chimeria: Musical Identity Social Network

In Chimeria: Musical Identity Social Network, the Chimeria Engine models category membership based upon musical preferences that are automatically constructed from a user's set of music “likes” (binary indications of positive valuation) on a social network profile. These “likes” constitute a set of musical artists from which we extrapolate, using commercially available musical classification data, moods (e.g., cheerful, gloomy, etc.), themes (e.g., adventure, rebellion, etc.) and styles (e.g., film score). This provides the context for non-binary group membership and passing (the “ability of a person to be regarded as a member of social groups other than his or her own...generally with the purpose of gaining social acceptance,” Renfrow, 2004). Each user's set of moods, themes and styles, then impacts the generated narrative in fundamental ways. We construct a conversational narrative on a social network structured by a model of conversation from sociolinguistics (Polanyi, 1989).

The Chimeria Application Interface consists of a procedurally generated photowall: a dynamic collage of photos representing the user's musical taste preferences. A feed of recent updates, posts, and invitations appear in an adjacent vertical timeline (see Figure 2). The system reacts to the user by generating interaction events from computer-controlled users who make up the user's social circle within the system.

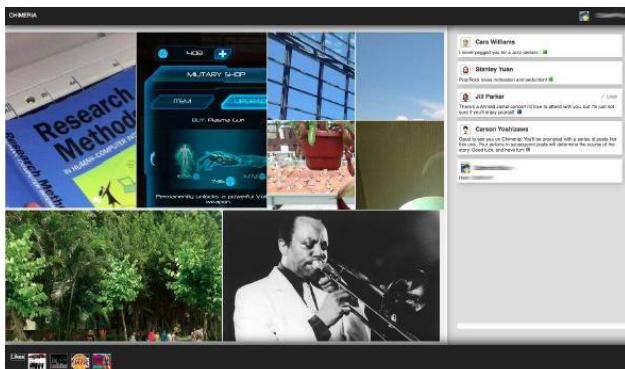


Fig. 2: A screenshot of the Chimeria: Musical Identity Social Network application interface

Figure 3 presents a screenshot of Chimeria: Musical Identity Social Network. Using musical preferences from the user's Facebook music likes or by manual entry, a hybrid real/fictitious narrative experience progresses over time. A series of dynamically generated posts by the user's friends (non-player

characters) comment on the user's membership within one or more musical affinity groups (i.e. “You're a raucous rock fan now?” or “Want to hear some airy jazz music?”). The user may “like,” “dislike,” or simply ignore these posts, resulting in group membership changes illustrated by alterations to a self-updating “photowall.” Some friends might question newly discovered interests, while others might pass judgment on prior affiliations. The resulting narrative may describe passing or assimilating as a member of a new group, reinforcing a prior group affiliation, or even being marginalized in every group. Some groups are deemed oppositional, privileged, or marginalized relative to others.

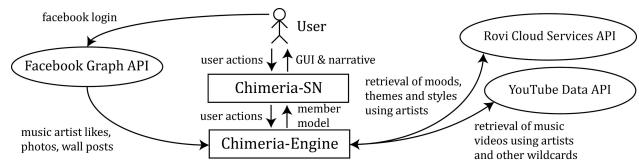


Fig. 3: The Chimeria Platform Applied to Musical Identities in a Social Network

3.2 Chimeria: Gatekeeper

Chimeria: Gatekeeper models a common RPG scenario – a player trying to gain access to the inside of a castle. Within this sample application, we demonstrate the power of the Chimeria Engine for enhancing this scenario by modeling more complex, adaptive, and nuanced conversations between PCs and NPCs, overcoming limitations identified in other videogames (Harrell et al., 2014). Figure 4 shows a preliminary visual design from Chimeria: Gatekeeper.

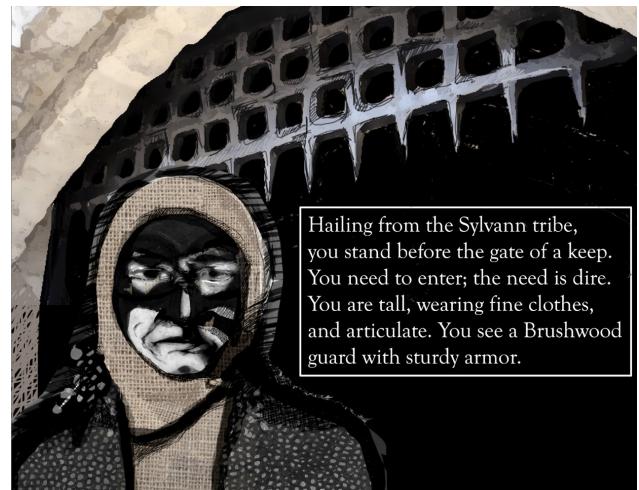


Fig. 4: Chimeria: Gatekeeper preliminary scenario visual design

Drawing on the work of Erving Goffman (Goffman, 1963), Chimeria: Gatekeeper attempts to model the effect of stigma on conversation. Within the scenario, the PC is initialized to the ‘discredited’ category and the NPC to the ‘accepted’ category. The accepted category is prototypically defined as the Brushwoods race – short, plain-spoken, and wearers of rough spun clothing. The discredited category is prototypically defined as the Sylvanns race – tall, well-spoken, and wearers of fine clothing.² To gain access to the inside of the keep, the player has to convince the guard that she or he is among the accepted category, in effect “passing” as a member of the category that has been instantiated as “accepted” (Harrell et al., 2014).

User actions and responses (e.g., slouching to adopt the posture of a prototypical Brushwood or displaying fine Sylvann clothing) incrementally shift the NPCs model of the PCs membership with respect to the categories, bringing the player closer to gaining access to the keep or to being rejected. Internal thoughts of the PC emphasize trade-offs between gaining utilitarian access to the keep and the loss of self-identity that can occur in trying to pass. The guard's responses of

approval or disapproval respond accordingly to chosen actions. A transcript of a run-through of Chimeria: Gatekeeper is shown in Figure 5.

Using Goffman's notion of impression management, we handle alternatives to the common trajectory of intentionally passing by considering other player decisions such as voluntary disclosure of stigma and slipping (trying to pass as a member of an accepted category, but failing). The modeling of passing and social categorization membership in Chimeria: Gatekeeper seeks to capture the stakes and power relationships often at play in real world social interactions.

The Sylvann and the Brushwoods have been at war for ages. The Sylvann, known as a tall people on average, are sometimes judged from afar to be lovers of finery and elaborate poetry. The Brushwoods, known as small people on average, are sometimes judged from afar to be fond of earthy homespun fabrics and good hearth tales.

Hailing from the Sylvann tribe, you stand before the gate of a keep. You need to enter; the need is dire. You are tall, wearing fine clothes, and articulate. You see a Brushwood guard with sturdy armor.

The Guard before you looks preoccupied.

The Guard is looking away from you:

The Guard smiles approvingly.

I'm trying to fit in with these Brushwoods.

A Guard stands before you, ready to size you up.

The Guard stares at you:

The Guard approves.

I'm trying to fit in with these Brushwoods.

The Guard before you has a wary expression.

The Guard asks you a question: "We don't see many, um, ...new...folk around these parts. You travel far to get here!" (It seems that he was about to say "many Sylvann.":)

"Tis not far from home."

"New? I'm from just around the way."

The Guard frowns disapprovingly.

I think he'll still let me in, though maybe I didn't give him the response he was looking for.

The Guard before you looks curious.

The Guard looks expectantly at you:

"Seasu re a' lle." (Sylvann for "Pleasant day to you.")

"Good day!"

The Guard smiles approvingly.

I'm trying to fit in with these Brushwoods.

Guard: "Welcome, I think you'll be at home here."

"Well, I got in, though I had to pretend to be something I'm not."

Fig. 5: Chimeria: Gatekeeper sample run-through

4. Conclusion

In this proposal, we have presented Chimeria, a platform for creating and analyzing narratives related to social group membership. By modeling character identities in a dynamic and nuanced fashion, we explore complex identity phenomena. By modeling social identity phenomena related to categorization, we use Chimeria to suggest how to better critically examine and express how identities are negotiated using digital media systems.

References

1. We have additionally implemented a small demo showing the applicability of Chimeria to a 3D game interface built in Unity.
2. We implement both "abstract" categories such as 'accepted' or 'discredited' and "concrete" categories that can instantiate them. This enables a great degree of flexibility for changing social dynamics within different contexts in games or different games altogether.

Bowker, G. C., & Star, S. L. (1999). *Sorting Things Out: Classification and Its Consequences*. Cambridge, MA: MIT Press.

Goffman, E. (1963) *Stigma: Notes on the Management of Spoiled Identity*. New York, NY: Simon and Schuster.

Harrell, D. F. (2010). *Toward a Theory of Critical Computing: The Case of Social Identity Representation in Digital Media Applications*. CTheory. <http://www.ctheory.net/articles.aspx?id=641> (accessed March 3, 2014).

Harrell, D. F. (2013) *Phantasmal Media: An Approach to Imagination, Computation, and Expression*. Cambridge, MA: MIT Press.

Harrell, D. F., Kao, D., & Lim, C. U. (2013). *Computationally Modeling Narratives of Social Group Membership with the Chimeria System*. In Proceedings of the 2013 Workshop on Computational Models of Narrative. pp. 123-128.

Harrell, D. F., Kao, D., & Lim, C. U., Lipshin, J., Sutherland, A., Makivic, J., and Olson, D. (2014). "Authoring Conversational Narratives in Games with the Chimeria Platform." In Proceedings of Foundations of Digital Games 2014. Ft. Lauderdale, FL, USA, Apr 3 – Apr 7.

Lakoff, G. (1987). *Women, Fire, and Dangerous Things*. University of Chicago Press, Chicago.

Linde, C. (1993). *Life Stories: The Creation of Coherence*. Oxford, U.K.:Oxford University Press.

Liu, H. (2007). *Social Network Profiles as Taste Performances*. Journal of Computer-Mediated Communication, 13: 252–275.

Polanyi, L. (1989). *Telling the American Story: A Structural and Cultural Analysis of Conversational Storytelling*. Cambridge, MA: MIT Press.

Renfrow, D. G. (2004). *A cartography of passing in everyday life*. Symbolic Interaction 27(4):485-506.

Framework of an Advisory Message Board for Women Victims after Disasters

Hashimoto, Takako

Chiba University of Commerce, Japan

Shirota, Yukari

yukari.shirota@gakushuin.ac.jp

Gakushuin University, Japan

1. Introduction

After the East Japan Great Earthquake, women victims have been suffering from different problems and worries: they had to care elders, raise children, and find jobs. They needed women-specific items. Administrative authorities wanted to recognize women victims' specific problems and give them appropriate supports. However, it was difficult to grasp women victims' requirements timely, because they were really patient and their needs were sometimes neglected under the environments and conditions that had changed from moment to moment. Because it takes a lot of labor to conduct interviews or questionnaire investigations to acquire their needs, we need a more useful and easily way to obtain women victims' needs.

To properly detect and analyze needs of women victims, we set our research goal was a development of an advisory message board for women victims on the web. To achieve our goal, this paper aims to make clear issues for acquiring women victims' needs and propose the frame work of an advisory message board for women victims after disasters. In our proposed message board, women victims can post their messages freely, and data mining techniques are utilized to analyze messages and detect specific needs.

2. Our Previous Work

We have already developed the prototype system for detecting time series victims' needs changes from social media data. The target data was the social media provided by the

non-profit organization that aims the March 11 earthquake and Tsunami relief. In our previous work, time series victims' needs changes/transitions were shown as changes of topics by adopting data mining techniques (Figure 1).

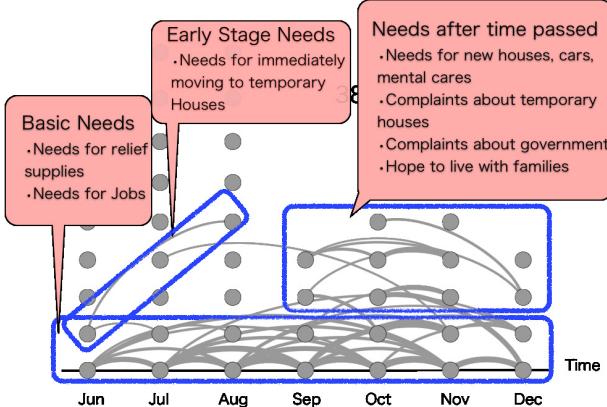


Fig. 1: Example of our previous work result (小文字にすべき単語が)

In Figure 1, the horizontal line shows time, and each circle shows one requirement. Lines between circles show transitions between requirements. There were three types of needs in affected people concerning the time length. The first is basic needs that appear for a long time, the second is early stage needs that appear immediately after the earthquake, and the third is needs after time passed. Needs about relief supplies and jobs are recognized as basic needs. On the other hand, there are needs for moves to temporary houses and evacuation center improvements at an early stage. After time passed, needs have changed, for example needs for new houses, cars, and mental cares, complaints about facilities of temporary houses, fears about the uncertain future, and wishes of living with families, appear.

Our previous method is useful to visualize victims' needs changes/transitions easily. However, the extracted needs were basically common to all. They were not special to women victims. Actually, most writers of the blog are male, and women victims tend to hesitate to publish their opinions, needs, or problems on social media.

3. Developing Framework of an Advisory Message Board for Women Victims after Disasters

3.1 Target Data

As a source of data, we used the following data on the web:

- Case Study Data: "The Support Women Victim Wanted! A Collection of Good Practices in Disaster Responses based on the East Japan Disaster"

This booklet collects examples of disaster response activities undertaken by various organizations in, and after, the East Japan Disaster. The data was collected by interviews and questionnaires and provided by the non-profit organization, Women's Network for East Japan Disaster(Rise Together).

To design a framework of an advisory message board, it is important to understand what kind of requirements women victims have, and make clear issues for their needs acquisition.

3.2 Issues on Acquiring Women Victims' Needs

For the target data[9], we adopted the morphological analysis technique to extract keywords for characterizing women victims' needs. Then, we evaluated each keyword and tried to connect it to one of three types of needs extracted by our previous method (Table I).

Table I. Relationships between general needs and women victims' specific keywords

General Needs (men and women)	Women Victims Specific Needs
Basic Needs Needs for relief supplies Needs for Jobs	Sanitary goods, Shorts, Bladder control pads, Cosmetics, Sunscreen, Babys' diaper, Burglar alarm, etc. Day nursery, Day-care center for elders, Incubation
Early Stage Needs Needs for immediately moving to temporary Houses	Female Workers, Clear temporally toilet, Women Area, Respite days
Needs after time passed Needs for new houses, cars, mental cares Complaints about temporary houses Complaints about government Hope to live with families	Networking among women, Health consultation Women area, Gender-segregated toilets, European style toilet Female workers, Violence hotline, Translation for foreigners Support for caring elders, support handicapped people

For example, the relief supplies that women victims needed were "sanitary goods", "shorts", "bladder control pads", "cosmetics", "diaper", "burglar alarm", etc. As for the needs of jobs, women victims needed "day nursery", "day-care center for elders", and "incubation". As for the complaints about governmental responses, women victims needed "female agents", "a violence hotline", and "a translation for foreigners" at the early stage. As for the complaints about temporary houses, women victims needed women areas, gender-segregated toilets, and so on. As for the needs of mental care, "networking among women", and "health consultation" were requested as women victims' specific needs.

Women victims needs did not appear in the general social media data. We found analyzing just general social media data was not enough for acquiring women victims' needs. Issues on acquiring women victims' needs are as follows:

- Women tend to hesitate to unfold their specific needs
- Women's needs tend to be put little emphasis on
- An interview and a questionnaire take a lot of labor
- Because after the disaster, women become very busy, it is difficult to make a time to post messages

3.3 Framework of an Advisory Message Board for Women Victims after Disasters

Figure 2 illustrates our proposed framework.

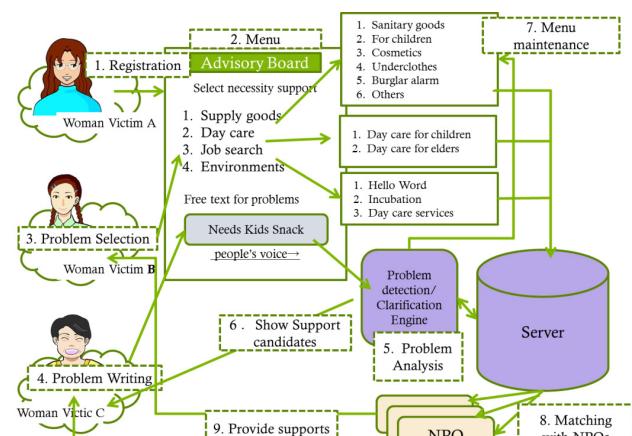


Fig. 2: Proposed framework: an advisory message board for women victims after disaster

We suppose that the board is organized by a responsible organization such as a public agency. In our framework, first, women victims who want to use this service, register themselves (1). Basically, it is an anonymous board but only

administrator can identify them. "Anonymity" is the point for acquiring women victims' needs easily.

Then they will select their needs/problems/requirements from the menu provided by the system (2, 3). The menu is hierarchically designed in advance according to our investigation. By providing problems candidates from the menu, the time for inputting messages should be decreased because they are busy. Then, solutions for solving their needs will be shown to them. If there are no corresponding problems in the menu, victims can input a text message freely. As for the free text, data mining technique will be adopted for analyzing problems, and solution candidates will be shown to them (5, 6). The menu contents will be maintained according to victims' inputs and responses.

The points of the framework is to clarify and predict problems of women victims, and continue to be improved according to their inputs and responses for providing appropriate supports at appropriate timing.

4. Conclusion

In this paper, first, we introduced our previous work, and then made clear issues for acquiring women victims' needs. Finally we proposed the frame work of an advisory message board for women victims after disaster. In our proposed message board, women victims can post their messages (requirements/opinions/complaints) freely, and data mining techniques are utilized to analyze messages and detect specific needs.

As the future work, we will develop the prototype system of the framework, and conduct the field work with non-profit organizations and actual women victims to receive feedbacks from them.

References

- Bannya Nippou, sviwatebanya.wordpress.com
 SAVE IWATE, <http://sviwate.wordpress.com/in-english/>
- Hashimoto, T., Kuboyama, T., Chakraborty, B., Shiota, Y.**, (2012), *Discovering Topic Transition about the East Japan Great Earthquake in Dynamic Social Media*, Proc. Of GHTC 2012, 259–264.
- Hashimoto, T. and Chakraborty, B.**, (2013) *Temporal Awareness of Needs after East Japan Great Earthquake using Latent Semantic Analysis*, Proc. of EJC2013, 214-226.
- Newman, M. E. J.** (2006), *Modularity and community structure in networks*, National Academy of Science USA 103(23), 8577–8696..
- Landauer, T. K., Dumais, S. T.** (1997), *A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge*, Psychological Review, 104(2), 211–240.
- Salton G, McGill MJ.** (1986), *Introduction to modern information retrieval*. McGraw-Hill. ISBN 0-07-054484-0.
- Church, K. W., Gale, W. A.** (1995), *Poisson mixtures*, Natural Language Engineering 1, 163– 190.
- Integrating Gender and Diversity Perspectives into Disaster Response: The Support We Wanted! A Collection of Good Practice in Disaster Response based on the East Japan Disaster*, risetogetherjp.org/?p=3909
- Women's Network for East Japan Disaster (Rise Together)*, risetogetherjp.org

Quelle médiation numérique pour le patrimoine bâti ?

Hennebert, Jérôme
jerome.hennebert@hotmail.fr
 Université de Lille 3, France

Résumé :

Dans un précédent projet de recherche et dans le contexte des musées de beaux-arts, nous avons échafaudé l'hypothèse selon laquelle la conception des audioguides de nouvelle génération, appelés « visioguide » (sur Iphone), ne répondait pas assez à la demande du visiteur, lequel souhaite d'abord augmenter son expérience esthétique (H2PTM, 2013). Nous proposons ici un nouveau modèle d'hypertextualisation des applications numériques destinées à la visite du patrimoine bâti, sur tablette de type Ipad, en suivant le modèle de l'abduction proposé par C. S. Peirce. Pour cela, nous comparons deux applications numériques qui correspondent à deux modèles opposés de médiation : par distanciation et par immersion 3d. Quel monde virtuel créer pour une médiation du patrimoine architectural par le numérique ? Quelles compétences sémiotiques l'usager doit-il mobiliser ?

Dans cette communication, nous approfondissons notre recherche sur la médiation apportée par le numérique pour la découverte touristique du patrimoine bâti : visite in situ ou de chez soi pour préparer la visite d'un monument. Dans la perspective des SIC, nous nous posons plusieurs questions : quelles informations architecturales transmettre dans une application pour tablette ? Les commentaires limitent-ils ou enrichissent-ils ce que l'architecte Rasmussen a appelé « l'expérience architecturale » (Rasmussen, 1959) ? Selon l'architecte danois, celle-ci ne peut être réduite à l'observation de plans ou de maquettes, puisqu'il ne s'agit pas tant d'expliquer que d'éprouver l'espace architectural et d'interpréter ses signes.

Notre hypothèse est de renouveler l'approche du « document monumental » (Patrick Fraysse, 2012) en relayant un discours sur l'œuvre bâtie par un discours entre l'œuvre et le public visiteur. Autrement dit, comment mieux restituer le lien intime entre un patrimoine bâti et la vie des hommes via une application numérique interactive ? Nous proposons dans cette communication un modèle innovant pour combiner les enjeux de la médiation par le numérique tantôt centrée sur les propriétés du patrimoine (Pantheon Iview) tantôt sur l'expérience du visiteur (Paris 3d saga).

Nous comparons deux applications numériques répondant à deux modèles opposés de médiation (Flon, 2012) : la « distanciation » avec Pantheon Iview (audioguidage didactique) ou « l'immersion » dans Paris 3Dsaga. Comment ces applications numériques rendent-elles la complexité d'une architecture ? La première application enrichit les connaissances historiques sur le Panthéon de Rome (Pantheon Iview). La seconde présente la ville de Paris comme un « monde » reconstitué à travers les âges (Paris 3dsaga par Dassault systèmes, 2012).

Les théoriciens de l'école de Constance (Jauss, Iser), ainsi que les théoriciens de la sémiotique (Eco, Mucchielli), ont montré que le sens n'est pas une donnée de départ inhérente à l'œuvre, mais co-construite par le récepteur : comment faire en sorte que le spectateur participe à l'élaboration du sens d'un édifice ? Comment une application numérique constitue-t-elle un interprétant fiable du patrimoine ? Comment donner au visiteur d'un bâtiment des compétences sémiotiques qu'il n'a pas reçues par les programmes éducatifs (Segaud, 1994) ?

Afin de répondre à ces questions, nous nous référerons à la conception du signe architectural selon Eco, comme « signe-fonction »

1. La communication architecturale est d'abord une stimulation, un complexe de sensations. Ex. : la vue de l'escalier me dispose à monter / descendre, donc me dispose à un usage fonctionnel ; la forme architecturale (signifiant) rend possible la fonction (signifié).
2. Les signes architecturaux sont ensuite de l'ordre de la perception ; ils risquent de ne pas être interprétés à cause de nos habitudes (pour monter, nous oublions facilement la fonction de l'escalier) ou par méconnaissance du code architectural, d'où perte du sens.

Partant de ces postulats, les applications numériques analysées devraient :

- a. décrire des signifiants qui à la fois dénotent les fonctions de l'architecture mais surtout connotent des idéologies (une conception de l'habitation à un moment donné). Ex : la

- courbe Art nouveau ou l'ogive gothique n'ont pas la même connotation pour déterminer l'objet fonctionnel « fenêtre ».
- b. Evoquer le signe-fonction à partir de codes architecturaux qui appartiennent au passé (ex. : l'Art nouveau est un code contre la standardisation industrielle fin XIXe s. mais pour les thèses socialistes et progressistes.)

Force est de constater dans les deux applications que le jeu des connotations et l'expression du code architectural font défaut. La simulation architecturale par immersion 3d est bien une stimulation (un ensemble de sensations) dans Paris 3d saga mais il manque des perceptions fines des signes architecturaux. L'approche y est plus historique et synthétique qu'analytique (seulement les étapes de construction de la cathédrale Notre-Dame de Paris ou de la Tour Eiffel).

Autrement dit, la médiation par le numérique surdétermine le rapport au signifiant visuel (réalité virtuelle dans Paris 3d saga), au détriment du rapport connotatif entre le signifiant et le signifié. Les applications focalisent ensuite sur le référent (la construction dans Paris 3D saga, et le contexte historique dans tout le corpus), au détriment de la signification architecturale. Il s'agit moins d'une médiation semio-cognitive que d'une « oocularisation » stimulante de l'architecture – une médiatisation technologique. Certes, la médiation vise ici à renforcer les connaissances historiques de l'usager par des commentaires contextuels et hypertextuels mais sans visée fonctionnelle. Au final, comprendre est ici plus apprendre l'histoire qu'explorer ou « éprouver » les signes architecturaux au sens de Rasmussen.

Par ailleurs, l'architecture, comme l'ont montré Eco et Rasmussen, est mal appréhendée lorsque :

1. le visiteur n'est pas relié directement à l'espace sensible d'un édifice
2. lorsque l'analyse du signe architectural n'est pas orientée vers un futur (une étude diachronique des signes

Le premier point est résolu grâce l'immersion 3D et au « gyroscope » dans Paris 3d saga. Le deuxième point reste problématique. Une œuvre d'art architecturale modifie les codes de son époque et programme une évolution des manières d'habiter. Cet aspect est négligé dans les deux applications. Paris 3d saga est une machine à remonter le temps, un feuilleté d'époques successives qui ne montre pas les différences de réception et d'usage d'un monument à travers différentes époques (la fonction « photogommage » dans Paris 3d saga est plus passéiste que visionnaire). Il n'est donc pas aisément pour le visiteur en quête de sens de se poser des questions sur l'architecture en phase d'immersion 3d, sans objectifs de recherche.

Repositionnons dès lors la médiation par le numérique comme une « mise en intrigue » du patrimoine et une « refiguration » du monde sensible du visiteur (Ricoeur, 1985). La refiguration est le troisième temps de la représentation pour le phénoménologue (« mimèsis3 »). Rappelons que le philosophe distingue :

1. Le temps de l'action (l'histoire) = mimèsis 1
2. Le temps de la narration de l'action (récit) = mimesis 2
3. Le temps du lecteur qui renvoie le récit à son propre « monde » et à sa sensibilité = mimèsis 3

Dès lors, une médiation par le numérique autre que celle proposée actuellement dans les processus d'audioguidage devient possible : une visite certes immersive comme dans Paris 3d saga, mais en plongeant l'usager dans un récit-enquête qui serait autant l'interprétant (mimèsis 2) qu'une source de plaisir (mimèsis 3). Un édifice constituant une énigme à résoudre (ex. la fiction possible d'une reconstitution), la médiation patrimoniale devient alors un récit exploratoire. Mais avec quel protagoniste ?

Afin de mieux articuler le signifié et l'éprouvé, afin de mettre en récit le patrimoine bâti, nous proposons de croiser deux modèles de médiation identifiés dans notre corpus :

1. Pour structurer les informations architecturales, nous proposons d'appliquer le modèle de l'abduction proposé par Peirce : un questionnement « impliquant » au cours d'une enquête herméneutique (méthode du détective pour déchiffrer les signes), à partir d'indices, et selon trois niveaux d'expérimentation :

1. a. le sensoriel ou l'esthésie (Boutaud, 2007) : la perception de l'espace-temps architectural par l'immersion 3d.
1. b. le sensible ou l'esthétique : perception du rythme des ouvertures en façade etc.
1. c. le relationnel ou l'éthique : établir le lien entre le sujet et le bâtiment d'une part, voire la culture et le collectif d'autre part.
2. En raison de la double performativité qui caractérise les applications (dire l'architecture et organiser une visite), nous distinguons le faire-savoir architectural du « faire-visiter » (organiser la visite). Pour ne pas dissocier les deux, nous proposons un modèle hybride par immersion et identification de l'usager au rôle de l'architecte (position réflexive et sensible).

C'est donc le projet de l'architecte qui constitue le grand oubli de ces applications. Si le visiteur se met à la place fictive de l'architecte, il pourra mieux éprouver le patrimoine bâti, en se posant des questions pertinentes sur le projet d'édification et son idéologie (lien de l'édifice avec son environnement physique et sa culture etc.).

En conclusion, la mise en récit du patrimoine et du projet de l'architecte favorise une refiguration possible de la réalité virtuelle, une médiation par le numérique différente, en faveur d'un apprentissage sensible de l'architecture. La médiation apportée par l'immersion 3d est relayée par un questionnement abductif et hypertextuel, au choix du visiteur selon sa culture architecturale. Nous retrouvons les principes de base d'un « serious game » pédagogique, aux fins d'une meilleure information sur l'architecture complexe et sur l'organisation même de la visite (en ajoutant la géolocalisation), enfin aux fins d'une meilleure communication entre visiteurs d'un groupe (famille, amis...) qui partageraient leur « expérience architecturale » (Rasmussen) autour d'une même fiction exploratoire.

Comme le suggère Jacques Rancière, aux fins politiques d'une meilleure démocratisation de la culture : « sur un monde, on ne fait pas de théorie, on fait son propre poème » (Ebguy, 2012). On peut donc imaginer de nouveaux aspects à la fois heuristiques et inventifs dans les applications numériques. Chaque récepteur devrait idéalement se réapproprier l'œuvre architecturale comme « vision personnelle du monde » à partir d'une scénarisation, et non comme une vérité historique imposée.

References

- Balpe J.-P.** (1990), *Hyperdocuments, hypertextes, hypermédias*, Paris, Eyrolles.
- Boutaud J.-J.** (2007), « *Du sens, des sens. Sémiotique, marketing et communication en terrain sensible* », Semen, n°23, 2007 [consulté le 15 mars 2013]. URL : <http://semen.revues.org/5011>
- Ebguy J.-D.** (2012), « *La mésentente : le philosophe (Jacques Rancière) et le poéticien (Gérard Genette)* », in Fabula-LhT, n° 10, « L'aventure poétique », décembre 2012, URL : www.fabula.org/lht/10/ebguy.html , [consulté le 04 janvier 2014].
- Eco Umberto** (1972), *La Structure absente, introduction à la recherche en sémiotique*, Paris, Mercure de France (1^{ère} éd. 1968).
- Flon Émilie** (2012), *Les Mises en scène du patrimoine : savoir, fiction et médiation*. Paris : Hermès/Lavoisier .
- Fraysse Patrick** (2012), « *Images du Moyen Âge dans la ville : l'inscription spatiale de médiévalité* » Communication et langage, mars 2012, n°171, p. 3-17
- Rasmussen Steen Eiler** (2002), *Découvrir l'architecture*, du Linteaù éd. (1^{ère} éd. 1959).
- Ricoeur Paul** (1985), *Temps et récit, tome III : Le temps raconté*, Paris, Le Seuil.
- Segaud Marion** (1994), « *Compétence esthétique et culturelle architecturale du français ordinaire* », in Figures architecturales, formes urbaines, actes du congrès de Genève de l'Association internationale de sémiotique de l'espace, Anthropos, Lausanne, p. 209-220.

Varano Sandro (2010), *Un espace de navigation hypermédia dédié au patrimoine culturel bâti*. Éditions Universitaires Européennes.

Open content production in museums. A discourse and critical analysis of the museum in the digital age

Hidalgo Urbaneja, María Isabel

mbelhu@gmail.com
Universidad de Málaga

1. Introduction

As widespread interest in the Open Data movement has grown in recent years some museums from all over the world have started to share and provide their digital content with Internet users. These new practices not only involve users' interest and traffic increase on museums websites but also magnify the institutional transparency which provokes a different conception of the museum authority. Then, we can find different ways of making available to explore and download high resolution photographs or exhibition catalogues as well as more complex procedures as open APIs development in order to allow users to create their own application based on museums data.

This paper seeks to establish the main issues related to museum discourses within the Open Data activities framework. Two case studies were chosen which construct their institutional and public digital identity on data openness. The Smithsonian Cooper-Hewitt Collection and the Rijksmuseum are well recognized by the professional museum community concerned with digital practices. Both museums had obtained important awards at the Museums and the Web 2013 conference as well as being cited on several blog posts by some of the most relevant museum bloggers.

As L. Manovich argues: 'The use of software re-configures most basic social and cultural practices and makes us rethink the concepts and theories we developed to describe them' (2013, 33). This assertion implies a re-configuration of the museum in epistemological terms provoked by digital practices. Taking into account the Foucaultian notion of discourse, it is necessary to analyse both the museum digital content and the institution and professional framework that arrange and define the discourse itself.

The progressive Open Data implementation by institutions - not only by museums- could be interpreted as a sign of democratization even if this openness does not represent a new role for museum institutions. However, as G. Lovink (2012, 49) states 'visibility and transparency are no longer signs of democratic openness but rather of administrative availability.' Then we could understand Open Data mechanisms in museums as just a public service which maintains the museum authority status according to digital age knowledge dissemination dynamics. Thus the question is whether open data is becoming merely a user service or something else?

In terms of open content access, users play an essential role receiving and reusing museum contents. Through open APIs they can elaborate computer mobile or tablet based applications. Eventually this option is oriented to amateurs, owners of specific informatics knowledge. The rise of amateurism and participatory culture argues for the author status dissolution. This is reflected in the reinterpretation of museum institutional authority. The museum is likely to maintain its status while granting more privileges to users.

2. Case Studies

In order to study further this question, two case studies were chosen whose relevant use of Open Data is creating new museum models on the Internet, although the data typology used by them is diverse as well as their public strategy. On one side, the New York Smithsonian Cooper-Hewitt Collection¹ has presented a website in beta mode that reflects the physical state of the museum which is being refurbished at the same time. On the other hand, Amsterdam's new Rijksmuseum website² has been launched at the same time as the reopening of the museum.

One of the most significant aspects is that both museums have published their database APIs on Github³ allowing users to access the museum collections data or metadata and develop API based applications. The Rijksmuseum digital identity is defined by the appeal of high resolution images of the artworks - which are currently being offered in its own website app Rijksstudio- while the Cooper-Hewitt Collection strength lies in the provisional and documentary nature of its data - which is composed by the objects raw data and metadata- eventually improved by the museum staff as well as wikipedians. The museum become 'human and fallible', just like the public. Moreover, the Smithsonian Cooper-Hewitt Collection website is making an exemplary use of open data including other features as biographical pages enhanced Wikipedia integration, public and open geographic identifiers or information concordances with other institutions.

The applications created by developers in both cases are a reflection of the museums digital strategy and conception. Clearly the Cooper-Hewitt collection apps -or visualizations- tend to be more experimental in contrast to the Rijksmuseum ones which also are similar to other museum applications - developed by the museums themselves- and are based on a definitive database. Likewise, Rijksmuseum apps play with the high quality of the artworks pictures rather than textual data on the Cooper-Hewitt Collection apps, where developers draw attention to data visualization or automated Tumblrrs.

Some of the Cooper-Hewitt collection applications best features are its digital work in progress such as Curatorial Poetry Tumblr⁴ , a stream of decontextualized descriptive texts pulled from museum collection meta-data, elaborated by the museum staff, or the visualization by Ruben Abad about the collection colour history⁵, have been documented on the Cooper-Hewitt Collection Blog. The process of developing apps acquires value by itself and this is reflected on the blog that documents the museum staff working activities. On the other hand, some of the Rijksmuseum apps, such as Faces of the Rijksmuseum⁶ , which uses facial recognition, or Rijsksify⁷ , that mix a music playlist with the collection, were developed during the TNW Kings of Code Hack Battle 2012 held in Amsterdam, confirming once again a definition of the amateur user profile who is interested on APIs usage. Referring to P. Gorgels from the Rijksmuseum, they identify several target groups: the culture snacker, art enthusiasts and professionals, then to which group does the amateur/hacker user belong?

Until the launch of the Live API on Github, the Rijksmuseum had shared on the museum website some of the apps developed with its OAI API⁸ although it is worth noting that the Cooper-Hewitt Collection is going further, publishing on its blog interviews or text written by the developer itself about the elaboration of API based apps. Definitely this action could be interpreted as a sign of openness recognizing the value of users' contributions. The use of Github, a collaborative revised hosting service for software development projects, reveals the museums interest on hackers/developers behaviour. Also the Github adoption by Rijksmuseum could be a symbol of openness and adaptation and make us to think whether it is the museum or the public who decide to introduce standards.

3. Conclusion

It is clearly evident that these Open Data museum practices can represent different levels and a wide conception of openness. Although turning back to the introductory points, is the museum institution really breaking their walls thanks to this new trend or they are just building a new ones? In fact, these

two cases reveal the existence of new boundaries, the terms and conditions webpage of the Rijsmuseum API, for example, shows how the museum control the use of its data. Therefore, should users ever obtain the same status as the museum in terms of authority? By extension, are the applications developed by users being valued differently regarding the museums ones? Clearly museums do, in the digital age they still are institutions that decide what is worthy and deserves their recognition. Maybe two case studies hardly can explain the substantial changes that Open Data is bringing about or whether the general public is understanding them as a sign of openness or even if museums are using it as a marketing strategy. However those substantial changes and their new discourses about openness are modelling the museum idea in the digital age.

References

- Cairns, S.** A museum collection that never ends? Cooper-Hewitt's new online collection *Museum Geek* <http://museumgeek.wordpress.com/2012/10/03/a-museum-collection-that-never-ends-cooper-hewitts-alpha-online-collection/> (accessed 6 March 2014)
- Chan, S., Walter, M., Shelly, K., Solas, N., Barnes, G.** *Cooper-Lewitt Labs*, <http://labs.cooperhewitt.org/> (accessed 21 February 2014)
- Barthes, R.** (1977). The death of the author [La mort de l'auteur] in *Image music text* translated by Heath, S. London: Fontana Press.
- Foucault, M.** (1972). *Archaeology of knowledge* [*Archéologie du savoir*] translated by A.M. Sheridan Smith, London: Routledge.
- Foucault, M.** (1970). *The order of things: an archaeology of the human sciences* [*Mots et les choses*]. London: Routledge.
- Gorgels, P.** (2013). Rijksstudio: Make Your Own Masterpiece! In *Museums and the Web* 2013, Proctor, N. and Cherry, R. (eds). Silver Spring, MD: Museums and the Web. Published January 28, 2013. <http://mw2013.museumsandtheweb.com/paper/rijksstudio-make-your-own-masterpiece/> (accessed 10 February 2014).
- Lovink, G.** (2012). *Networks without a cause. A critique of social media*, Cambridge: Polity Press.
- Manovich, L.** (2013). *Software takes command*, New York: Bloomsbury Academic.
- Parry, R.** (2013) The Trusted Artifice: Reconnecting with the Museum's Fictive Tradition Online, in Drotner, K. and Schrøder, K. C. (eds) *Museum Communication and Social Media: The Connected Museum* New York and Abingdon, Oxon.: Routledge, pp. 17-32.
- Parry, R.** (2010). *Museums in a digital age*, London: Routledge, 2010.
- Parry, R.** (2007). *Recoding the museum: digital heritage and the technologies of change*, London: Routledge.
- Ridge, M.** *Museums and the machine-processable web* <http://museumapi.pbworks.com/w/page/21933420/Museum%20APIs> (accessed 21 February 2014)
- Waterton, E., Smith, L. and Campbell, G.** (2006). The Utility of Discourse Analysis to Heritage Studies: The Burra Charter and Social Inclusion, *International Journal of Heritage Studies*, 12 :4, pp. 339-355. 10.1080/13527250600727000
- Wilson, R.J.** (2011). Behind the scenes of the museum website, *Museum Management and Curatorship*, 26 (4) London, Routledge, pp. 373-389. 10.1080/09647775.2011.603934
1. www.cooperhewitt.org/# (accessed 21 February 2014)
 2. www.rijksmuseum.nl/# (accessed 21 February 2014)
 3. archive.is/LYWKG#selection-839.0-839.34 (accessed 14 February 2014)
 4. curatorialpoetry.tumblr.com/ (accessed 21 February 2014)
 5. dataclimber.net/blog/2014/1/19/cooper-hewitts-collection-color-history (accessed 21 February 2014)
 6. weblab.ab-c.nl/rijksmuseum (accessed 21 February 2014)
 7. hackbattle.thenextweb.com/index.php/Riksify (accessed 21 February 2014)
 8. archive.is/LYWKG#selection-839.0-839.34 (accessed 14 February 2014)

CLARIN: Resources, Tools, and Services for Digital Humanities Research

Hinrichs, Erhard

Eberhard Karls University Tübingen, Germany

Krauwer, Steven

Utrecht University, The Netherlands

1. Introduction

CLARIN is the short name for the *Common Language Resources and Technology Infrastructure*. It aims at providing easy and sustainable access for scholars in the Humanities and Social Sciences (HSS) to digital language data and advanced tools to discover, explore, exploit, annotate, analyse or combine them, independent of where they are located. CLARIN is one of the research infrastructures that were selected for the European Research Infrastructures Roadmap by ESFRI, the European Strategy Forum on Research Infrastructures. The CLARIN Governance and Coordination body at the European level is CLARIN ERIC. An ERIC is a new type of international legal entity, established by the European Commission in 2009. Its members are governments or intergovernmental organisations.

CLARIN is in the process of building a networked federation of European data repositories, service centres and centres of expertise, with single sign-on access for all members of the academic community in all participating countries. Tools and data from different centres will be interoperable, so that data collections can be combined and tools from different sources can be chained to perform complex operations to support researchers in their work. The CLARIN infrastructure is still under construction, but a number of participating centres are already offering access services to data, tools and expertise. The purpose of the present paper is to give an overview of language resources, tools, and services that CLARIN presently offers.

2. Reference Data Sets

The federation of CLARIN centers offers a high number of reference data sets that are well-known and widely used in the scientific community. The CLARIN Center in Vienna offers the Austrian Academy Corpus, a very large collection of German texts and German literature covering the period of 1848 to 1989. The German reference corpus DeReKo, the largest linguistically motivated collection of contemporary German texts with more than 4.0 billion word tokens, is hosted by the CLARIN Center in Mannheim. The CLARIN Center in Berlin provides access to the German Text Archive, a digital collection of German-language printed works from around 1650 to 1900 as full text and as digital facsimile. The CLARIN Center in Sofia offers the Bulgarian Reference Corpus. CLARIN Center in Warsaw hosts the National Corpus of Polish, a reference corpus with more than fifteen hundred million words.

CLARIN centers offer extensive collections of spoken language. The CLARIN Center in Amsterdam is home to thousands of hours of audio material for Dutch, including more than 1000 hours of dialect recordings. The CLARIN Center in Munich specializes in digital corpora for contemporary German. The CLARIN Center in Sofia offers the Bulgarian Political and Journalistic Speech corpus.

CLARIN language resources are not restricted to the languages spoken in CLARIN member countries. The CLARIN Center in Nijmegen offers easy access to the DOBES Archive, which documents endangered languages around the world.

Another key language resource are high-quality lexica. The CLARIN Center in Tartu provides on-line access to a variety of lexica for Estonian <http://www.keelevaab.ee/>. The CLARIN center in Berlin is home to the Digitale Wörterbuch der deutschen Sprache (DWDS). The DWDS lexicon uses

extensive digital corpus collections to document the actual usage of German words and offers on-line access to all materials at www.dwds.de/. Apart from traditional lexica, CLARIN also offers access to lexical resources that model word meanings in terms of a network of lexical and conceptual relations. The CLARIN center federation currently hosts such word nets for Czech, Danish, Dutch, Estonian, Finnish, German, and Norwegian.

In addition to reference data sets, CLARIN provides access to an extensive set of metadata records. The Virtual Language Observatory (www.clarin.eu/vlo) currently contains more than 500.000 metadata records to language resources and tool. Faceted search and a visual map provide easy-to-use interfaces for HSS scholars to locate language resources and tools that match the needs in a particular research project.

2.2. Creation of New Resources

For new digital data sets, special care must be taken that such data creation efforts adhere to best practices or standards for text encoding whenever possible and follow a data management plan. HSS scholars often lack the necessary experience or access to data repositories to meet these expectations. The CLARIN-D User Guide [1] provides practical information on the use of standards for language resources and on following good practices in data creation.

3. Data Mining and Data Analysis

3.1. Query Tools and Federated Content Search

Since data sets available in electronic form are typically very large, CLARIN centers support HSS scholars by providing powerful and easy-to-use query tools for many of the resources described above. Access is greatly facilitated if such query tools are realized as web applications and thus available in any web browser. Two good examples of this kind are the web application for querying the German Text Archive and the MIMORE (<http://www.meertens.knaw.nl/mimore/search/tool>) tool, which enables researchers to investigate morphosyntactic variation in the Dutch dialects by searching three related databases with a common on-line search engine. The search results can be visualized on geographic maps and exported for statistical analysis.

In addition to query interfaces for individual resources, CLARIN offers a Federated Content Search (FCS) functionality that enables HSS scholars to construct a virtual corpus collection hosted by different CLARIN centers and to query this virtual corpus via a common search interface. Currently, nine CLARIN centers in Germany and in the Netherlands make more than 20 resources available to the linguistic researchers via the common interface of the CLARIN-D Federated Content Search (weblicht.sfs.uni-tuebingen.de/Aggregator), and this number is growing. The CLARIN Center at the University of Oslo also provides FCS functionality via the GLOSSA corpus query tool. [5]

3.2. Workflows for Data Annotations

Language data that are annotated with linguistic information can be searched with high accuracy for specific data patterns. The CLARIN Centers in Oslo, Prague, Tübingen and at the Dutch Language Union offer linguistically annotated corpora, so-called treebanks, for Czech, Dutch, German, and Norwegian with accompanying query tools.

If a collection of language resources does not contain sufficient linguistic information, for example if the word forms in a corpus have not been lemmatized, it is impossible to obtain meaningful word frequency distributions. Likewise, if an HSS scholar wants to search for all person names in a very large newspaper corpus in order obtain an overview of who is currently in the news, then the person names in such a corpus needs to be marked up. CLARIN offers support for

HSS scholars who need to add annotations of this kind. The web application WebLicht [2], hosted by the CLARIN Center in Tübingen, is a tool-suite for automatic annotation of text corpora. Linguistic tools such as tokenizers, part of speech taggers can be combined into custom processing chains. The resulting annotations can then be visualized in an appropriate way, such as in a table or tree format. Recently the WebLicht tool suite has been extended to spoken language. This can be achieved with the integration of the WebMaus tool provided by the CLARIN center in Munich. WebMaus takes as input an audio file and its transcription and automatically aligns the speech signal with its transcriptions. The WebLicht tool can then further annotate the transcriptions so that via the automatic alignment, a user can find the relevant portions of the speech signal for particular data patterns.

4. Data Visualization

Visualization tools that render the data analysis results in an easy-to-grasp fashion are particularly important if the data sets involved are very large. While CLARIN cannot provide a comprehensive suite of eHumanities visualization tools, it can already support HSS scholars with a number of helpful applications. [3] The CLARIN Center at the University of Copenhagen has developed a visualization tool for parallel inspection of word nets. CinaViz[6] is web application provided that offers geo-visualizations for tracking city names with particular linguistic features.

5. Data Sharing and Data Archiving

CLARIN also provides support for the sharing, publishing and archiving of the data sets. SimpleStore and OwnCloud solutions are available for collaborative work on the same data set. Many CLARIN data repositories offer archiving services for external resources and for finished data sets. For quality assurance, all CLARIN Centers are assessed by the CLARIN Assessment Committee, according to strictly defined technical requirements (see: <http://hdl.handle.net/1839/00-DOCS.CLARIN.EU-78>) and have to obtain the Data Seal of Approval[7] for their services.

6. Conclusion

Interoperability of language resources and tools in the federation of CLARIN Centers is ensured by adherence to TEI and ISO standards for text encoding, by the use of persistent identifiers as long-lasting references to digital language data as well as by the observance of common protocols: Shibboleth for user authentication and authorization, SRU/CQL for Federated Contents Search, and OAI-PMH for metadata harvesting.

Here we could describe only a subset of all CLARIN resources and tools. For comprehensive and up-to-date information we refer interested readers to the CLARIN homepage: www.clarin.eu

References

- Herold, A. and L. Lemnitzer**, eds. (2012). *CLARIN-D User Guide*. Available at: de.clarin.eu/en/language-resources/userguide.html.
- Hinrichs, E., M. Hinrichs & T. Zastrow** (2010). *WebLicht: Web-Based LRT Services for German*. In: Proceedings of the Systems Demonstrations at the 48th Annual Meeting of the Association for Computational Linguistics (ACL-2010). Uppsala, Schweden. pp. 25-29.
- Zastrow, T., E. Hinrichs, M. Hinrichs, and K. Beck** (2013). *Scientific Visualization for the Digital Humanities as CLARIN-D Web Applications*. Proceedings of Digital Humanities 2013, University of Nebraska.
- The ESFRI Roadmap contains five research infrastructures in the area of Social Sciences (CESSDA, European Social Survey, and SHARE) and Humanities (CLARIN and Dariah).

en.wikipedia.org/wiki/German_Reference_Corpus .
nkjp.pl/index.php?page=0&lang=1
www.mpi.nl/dobes
github.com/textlab
weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/
CiNaViz_-_Visualization_of_European_City_Names
datasealofapproval.org/en

Trading Consequences: A Case Study of Combining Text Mining & Visualisation to Facilitate Document Exploration

Hinrichs, Uta

uh3@st-andrews.ac.uk

SACHI, University of St. Andrews

Alex, Beatrice

balex@staffmail.ed.ac.uk

ILCC, School of Informatics, University of Edinburgh

Clifford, Jim

jim.clifford@usask.ca

Department of History, University of Saskatchewan

Quigley, Aaron

aquigley@st-andrews.ac.uk

SACHI, University of St. Andrews

1. Introduction

This paper reports on interdisciplinary work carried out as part of Trading Consequences¹, a two-year Digging into Data project². The focus of the project is to mine large quantities of historical documents, extract information on commodity trading in the nineteenth century British World and visualise the mined output in dynamic and interesting ways, thereby bringing archives alive in ways that authors of original documents would have never imagined. The Trading Consequences interface is aimed at historians studying commodities and their environmental consequences. Their studies have tended to focus on a manageable number of commodities (e.g. William Cronon's research on beef, lumber and wheat³). The Trading Consequences Project aims at identifying global trends in commodity trading for many different natural resources, raw materials or lightly processed goods by correlating information extracted for one commodity with that of others or showing all commodities relevant to particular locations and dates.

In this paper, we first present an overview of this collaborative project that involved environmental historians, text mining, database experts and visualization researchers. We then report on lessons learned from a workshop where we collected feedback from historians and geographers after they interacted with the interface prototype in a series of exercises. This feedback informed the further adaptation of the underlying technologies for historical research.

2. Trading Consequences

The Trading Consequences system encompasses three main technical components: a text mining system, a database and a web-based user interface with dynamic visualisations (Fig. 1). The data analysed using this system is comprised up of several nineteenth century British and Canadian text collections⁴. These sources amount to over 11 million pages and over 7 billion analysed word tokens.

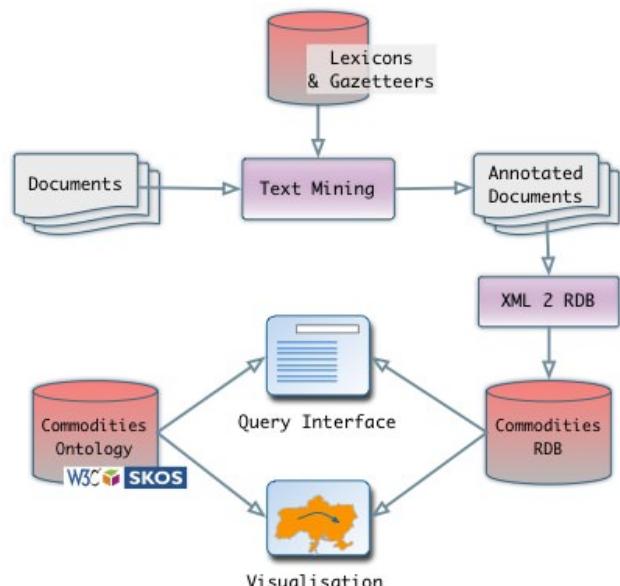


Fig. 1: System architecture.

2.1 Text Mining

The text mining (TM) tools are developed by the Language Technology Group at the University of Edinburgh. We adapted an existing pipeline built on LT-XML2 and LT-TT2 to process historic text⁵. The TM component is made up of a series of linguistic processing steps which build up the linguistic properties of the language in a given text. A pre-processing stage includes tokenisation, sentence-splitting, part-of-speech tagging and lemmatisation to determine words and sentences, identify their syntax and compute canonical forms of word tokens. All of this information aids down-stream TM processes. The next steps are named entity recognition and grounding. This means that mentions of locations, commodities and dates are automatically identified in the text and grounded to unique identifiers in existing knowledge databases. We ground location mentions to GeoNames identifiers and their corresponding latitude/longitude values⁶. We use an adapted version of the Edinburgh Geoparser for this geo-referencing process⁷. Commodity mentions are grounded to DBpedia⁸ concepts in a semi-automatically constructed commodity lexicon developed in this project⁹. Finally, date mentions are grounded to year, month and date attributes. The last TM step identifies relations between commodity, date and location mentions to identify the relevance of commodities in space and time. The extracted and enriched TM output is stored in a relational PostgreSQL database set up and hosted by EDINA¹⁰ for subsequent querying and visualisation.

2.2 Interactive Visualisations

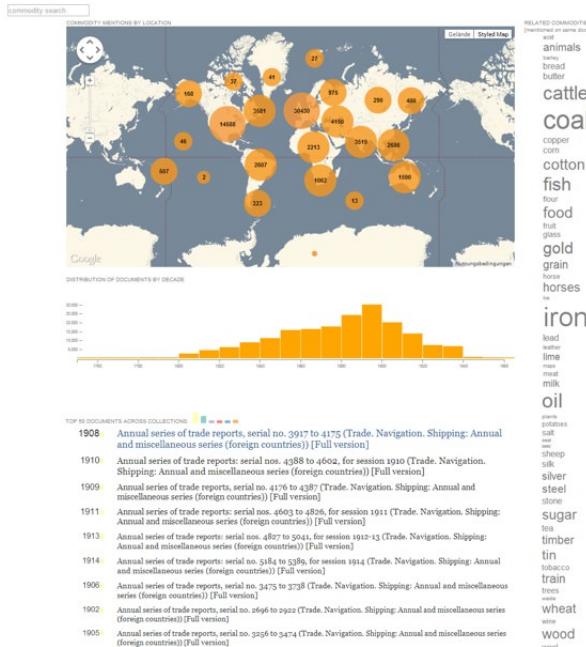
The strength of information visualisation is to make abstract concepts and relations within data visible and explorable¹¹. In the context of Trading Consequences we aim at providing visualisations of the mined data to:

1. - enable open-ended explorations of the document corpus beyond target search¹², i.e., supporting visual querying along spatial, temporal, and conceptual dimensions, and
2. - highlighting trends within a range of document data, for instance, relations between different commodity types.

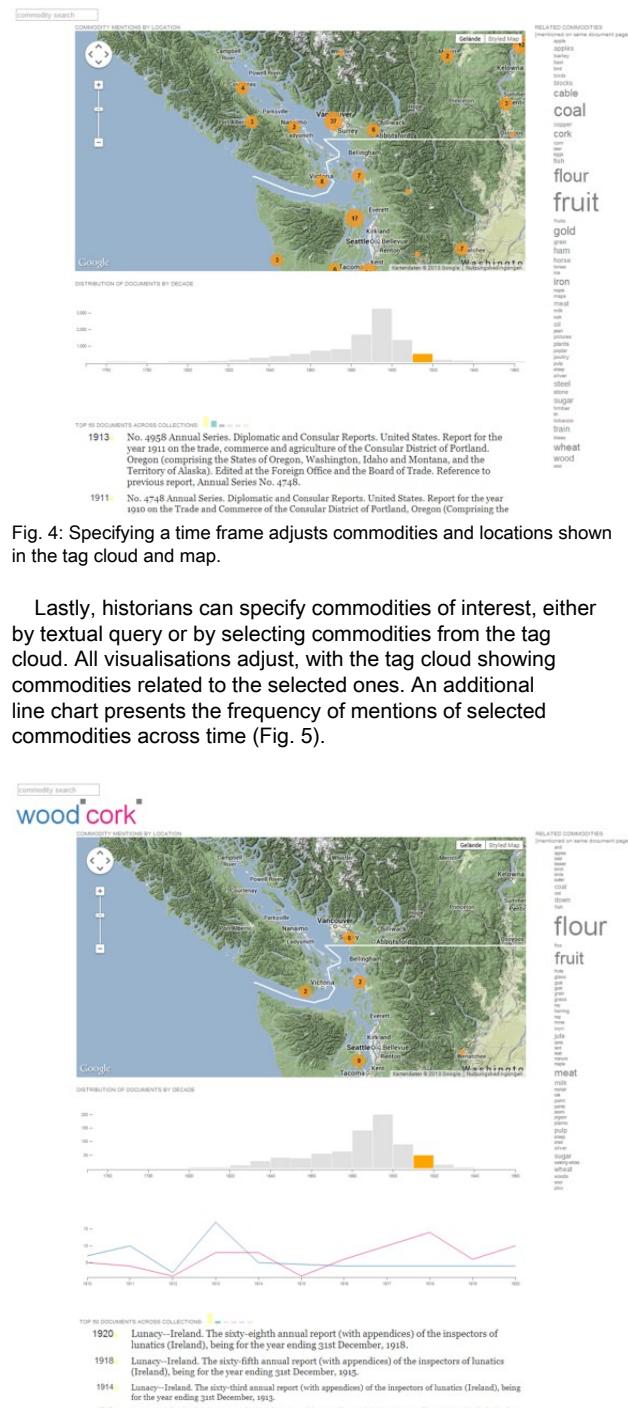
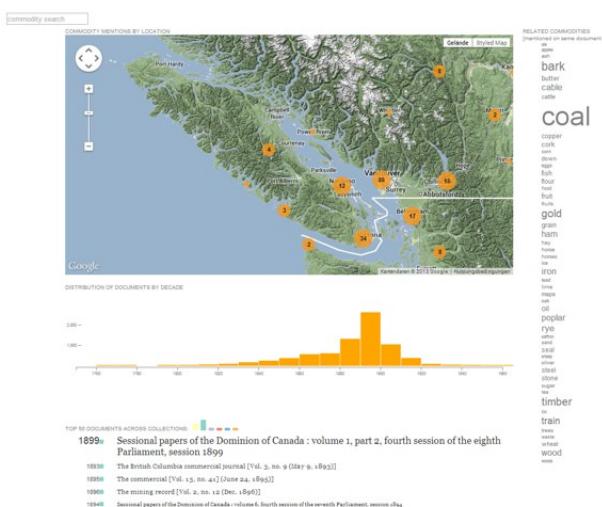
While the first approach facilitates the discovery of related documents in ways that common text-based search interfaces cannot, the second approach can lead to new insights or research questions based on collection sizes that exceed possibilities of traditional research methods in the humanities. Our visualizations are web-based to make them easily

accessible worldwide (see [1]). The implementation is based on JavaScript (D3.js and jQuery) and PHP.

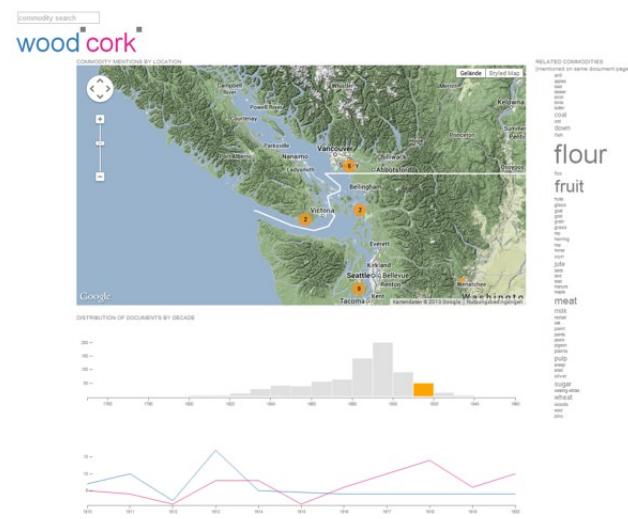
In this paper we briefly describe the Trading Consequences visualisation tool and how it was experienced by environmental historians as part of a workshop. Inspired by¹³, our visualisation consists of three interlinked representations (Fig. 2): a map showing the geographic context in which commodities were mentioned, a vertical tag cloud showing the 50 most frequently mentioned commodities, and a bar chart representing the temporal distribution of documents within the collection. A ranked document list provides direct access to the relevant articles.



Interaction with one visualisation acts as a filtering mechanism and adjusts the data shown in the other visualisations. For instance, zooming into the map adjusts the tag cloud to only include commodities mentioned in relation to visible locations; the bar chart only shows documents that include these commodity/location mentions (Fig. 3). Particular time frames can be selected to further filter the document corpus; the other visualisations are updated accordingly (Fig. 4).



Lastly, historians can specify commodities of interest, either by textual query or by selecting commodities from the tag cloud. All visualisations adjust, with the tag cloud showing commodities related to the selected ones. An additional line chart presents the frequency of mentions across time (Fig. 5).



3. Feedback from Historians

To gain expert feedback on our approach of combining text mining with visualisations to facilitate research in environmental history, we conducted a half-day workshop where we introduced our visualisation prototype to historians. The workshop was held at the Canadian History & Environment Summer School 2013 with over 20 environmental historians participating¹⁴. At the workshop, we asked historians to explore the visualisation tool in small groups (Fig. 6). To promote engagement with the different visualisations and to fuel discussions, the explorations were guided by a number of open-ended tasks, such as querying for commodities of interest or focusing on a geographic area.

Some historians immediately started to focus on the Vancouver Island area where the workshop took place. Others experimented with commodities and locations related to their

own research. In general, these first exploration periods were about verifying familiar facts to assess the capabilities of the visualisation and the trustworthiness of the underlying data. The historians quickly understood the general purpose and high-level functionality of the visualisations and were able to start their explorations immediately. There was some confusion, however, about lower level details. For instance, the meaning of the size and number of clusters in the map was unclear (e.g. do they represent number of documents, or number of commodity mentions?). Observing changes in the visualisations while adjusting parameters helped, but our observations highlight that clear labelling and tooltips are crucial for visualisations in the context of digital humanities, not only because these are a novel addition to traditional research methodologies, but also because they can be easily misinterpreted. The meaning of visual representations needs to be clear in order to make visualisations a valid research tool.



Fig. 6: User workshop at CHESS 2013.

Workshop participants found the meta-level overviews of the visualisations valuable as these can aggregate information about the document corpus beyond human capacity. In the short time of the workshop, historians made (sometimes surprising) discoveries that sparked their interest to conduct further research. While it is unclear if these discoveries withstand more detailed investigation (there is still some noise in the data), this shows that visualisation has the potential to support exploration and insight in the context of history research.

A large part of the discussions focussed on what kind of insights can be gathered from the visualisations. Some historians pointed out that the visualisations represent the rhetoric around commodity trading in the 19th century: they show where and when a dialogue about particular commodities took place, rather than providing information about the occurrence of commodities in certain locations. This raises the question of how we can clarify what kind of data the visualisations are based on to avoid misinterpretation.

4. Conclusion

In general, we received positive feedback about our approach of combining text mining and visualisation to help research processes in environmental history. Historians saw the largest potential in the amounts of data that can be considered for research but also in the open-ended character of the explorations that the visualisations support in contrast to common database search interfaces. Other types of visualisations were suggested to help analyse and discover relations and patterns in the data, something that we are currently developing.

Our future research will explore how our approach integrates into research processes in environmental history and how it can produce profound outcomes. This will involve controlled experiments including directed and open-ended tasks. We will also conduct long-term studies to evaluate the discoveries and limitations that historians encounter when using our tools. The wide-ranging feedback from the workshop was crucial in helping the computer science team members understand priorities and research methodologies of environmental historians. Expert feedback is an important component of interdisciplinary research in digital humanities.

References

1. Trading Consequences: tradingconsequences.blogs.edina.ac.uk/
2. Digging Into Data: www.diggingintodata.org/
3. **William Cronon** (1992). *Nature's Metropolis: Chicago and the Great West*. W.W. Norton, New York.
4. Data collections: tradingconsequences.blogs.edina.ac.uk/about/the-corpus/
5. LT-XML2: www.ltg.ed.ac.uk/software/ltxml2 ; LT-TTT2: www.ltg.ed.ac.uk/software/ltxml2
6. GeoNames: www.geonames.org/
7. **Claire Grover, Richard Tobin, Kate Byrne, Matthew Woollard, James Reid, Stuart Dunn, and Julian Ball** (2010). *Use of the Edinburgh Geoparser for georeferencing digitised historical collections*. Philosophical Transactions of the Royal Society A.
8. DBpedia: dbpedia.org. We accessed DBpedia via the SPARQL endpoint (dbpedia.org/OnlineAccess) , most recently on 16/12/2013, corresponding to DBpedia version 3.9.
9. **Ewan Klein, Beatrice Alex and Jim Clifford** (2014). *Bootstrapping a historical commodities lexicon with SKOS and DBpedia*, In: Proceedings of the LaTeCH 2014 workshop at EACL 2014.
10. EDINA: Jisc-designated centre for digital expertise & online service delivery; edina.ac.uk/
11. **Stuart K. Card, Jock D. Mackinlay, Ben Shneiderman** (eds.) (1999). *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann Publishers, Chapter 1, pp. 1-34.
12. **Gary Marchionini** (2006). *Exploratory search: From finding to understanding*. Communications of the ACM 49, 4, 41-46.
13. **Marian Dörk, Sheelagh Carpendale, Christopher Collins and Carey Williamson** (2008). *VisGets: Coordinated Visualizations for Web-based Information Exploration and Discovery*. IEEE Transactions on Visualization and Computer Graphics, 14(6), pp. 1205-1212.
14. CHESS 2013: 70.32.75.219/2013/04/12/cfp-canadian-history-and-environment-summer-school-2013-vancouver-island/

Tuning the Word Frequency List

Hoover, David L.

New York University, United States of America

One recent trend in computational stylistics and authorship attribution has been to “tune” the word list for better results. T-tests can identify words with statistically significant differences in frequency between two authors; they remain an excellent method for one-on-one problems (Burrows 1992; McKenna and Antonia 1996; Hoover 2010). Eder and Rybicki (2011) use elegant methods to identify “sweet spots” in the word frequency spectrum by testing a range of numbers of the most frequent words (MFW). They also progressively remove words from the top of the spectrum, eliminating the function words favored by so much previous work. Personal pronouns have been removed to minimize differences in point of view and the numbers of male and female characters (Hoover 2002). The word list has been “culled” by removing words that are frequent because of high frequencies in one text (Hoover 2003, 2004; Burrows 2005). Rybicki and Eder (2011), Jockers, Witten, and Criddle (2008), and Rybicki and Heydel (2013) have culled words absent from any or many texts. Burrows’s Zeta and Iota select words consistently used by one group and avoided by another, ignoring frequency. These tuning methods vary in effectiveness with the number, size, date, genre, and language of the texts, and with the number of authors involved.

I propose here a transparently motivated form of tuning with a strong a priori plausibility: selecting unevenly distributed words, using the coefficient of variation (CoV). This dispersion measure is defined as $S\text{tdev}/\text{Ave. Freq.} * 100$, expressed as a percentage of the average frequency. Consider the following four words, with statistics from 21 novels and short fiction by Willa Cather and Edith Wharton:

Word	Ave.	Stdev	CoV	# Texts	Rank
magnificent	0.0034	0.0064	192%	5	3374
longing	0.0034	0.0107	316%	5	2427
generally	0.0036	0.0059	163%	7	2994
father	0.0368	0.0600	163%	14	332

Magnificent and *longing* both appear in 5 texts with similar average frequencies but very different Stdev, CoV, and rank: *longing* varies much more in frequency than does *magnificent*. *Generally* and *father* show that the Stdev is too closely tied to the average frequency to measure dispersion here: low-frequency words tend to have small Stdev's. For the 21 texts above, among the 17,000 word types, only 1 of the 100 largest Stdev's comes from a word ranking 1,000 or higher. The frequencies and Stdev's of *father* are about 10 times those of *generally*, so both have the same CoV, which is thus a fairer measure of variability than the Stdev.

Unfortunately, simply analyzing words with the highest CoV is unworkable. The Stddev tends to be lower for less frequent words, but rare words and those occurring in a small number of texts have very large CoV's. For the 21 texts above, 9,000 of the 17,000 word types appear in only one text and share the largest CoV (447%), regardless of their frequency. Some character names in long texts rank as low as 63rd, but 7,200 are hapax legomena. Furthermore, only 300 of the words with the 12,000 largest CoV's appear in more than two texts. The need to identify words used fairly frequently and in many texts but with widely varying frequencies suggests combining Rybicki-style culling with the CoV in a method I call CoV Tuning. Now, some experiments. First, consider the Cather/Wharton set above, containing 3 Wharton novellas and 7 stories and 11 Cather stories. Standard cluster analysis (Ward linkage, squared Euclidean distance, standardized variables) correctly groups all texts in analyses based on 990 and 900MFW and fail at 400MFW only for Wharton's "The Hermit and the Wild Woman." (The seemingly peculiar choice of 990MFW arises from a limitation in my statistics program, *Minitab*.) All other analyses based on the 100-800MFW show at least 3 errors, including that same story.

I tuned the word list by sorting the 1,500MFW on the CoV. The words with largest CoV's, mainly character names, appear in only one text, so I re-sorted the words on the number of texts they appear in, retained only words appearing in 7 or more texts, then re-sorted on the CoV and retained the 1,000 most variable words (MVW). Cluster analysis using CoV Tuning is very effective: all 7 analyses based on the 400-990MVW correctly group all the texts. Retaining words appearing in at least 5, 6, or 8 texts is slightly less effective. T-testing the 1,500MFW and retaining only the 352 words with $p < .05$ gives perfect groupings using all 352 words and decreasing numbers of them, down to the two words with the highest T value, the classic authorship pair, *till* and *until*. Thus, for two-author problems, the t-test is clearly superior. Figures 1-2 show standard and CoV Tuned analyses.

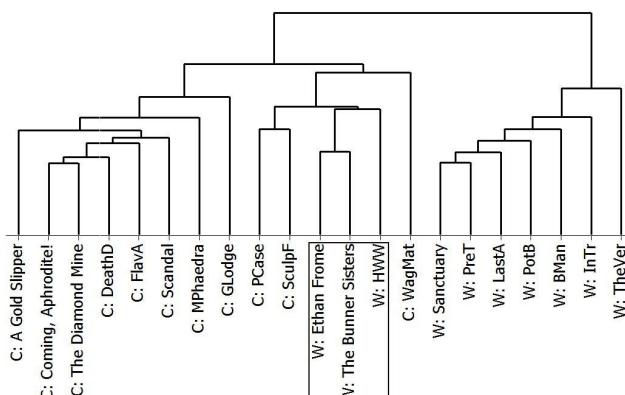


Fig. 1: Standard 800MEW

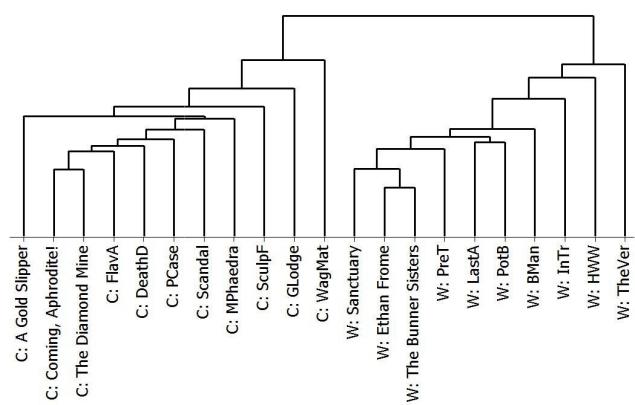


Fig. 2: CoV Tuned, 800MVW

Now consider some difficult multi-author problems for which t-test tuning is unavailable. First I tested a set of 43 late 19th and early 20th century novels by 15 American authors, 2-3 novels each. Errors for at least 3 authors occur in all analyses using the standard method. Using CoV Tuning retaining only words found in 34 or more texts, analyses of the 900 and 700MVW have just 1 error, and analyses of the 500, 600, 800, and 990MVW have 2. Minimums of 22 and 38 texts give weaker results. Figures 3-4 show standard and CoV Tuned analyses, respectively.

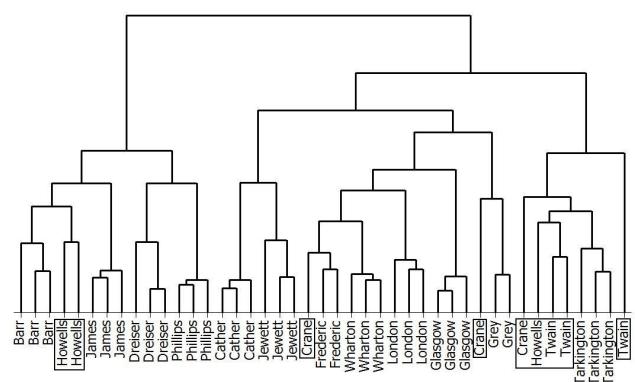


Fig. 3: Standard, 700MFW

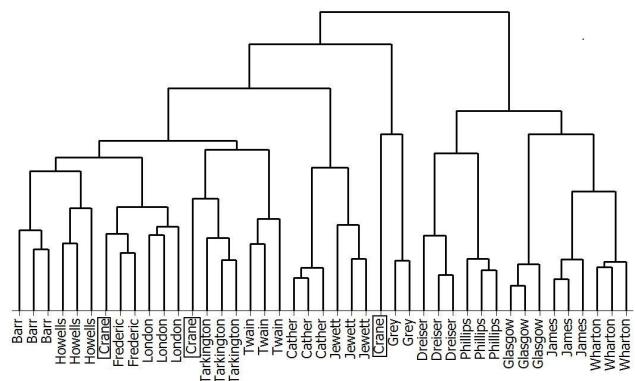


Fig. 4: CoV Tuned, 700MVW

Turning to a different genre and shorter texts, I tested 25 pieces of literary criticism by 14 authors, 9 authors with 2 texts each, 1 with 3 texts, and 4 with 1 text (see Hoover 2001 for details). I made this problem more difficult by dividing the texts into 33 sections of about 4,000 words. The standard method works well here, correctly grouping all sections by 12 of the 14 authors in analyses of the 600, 700, 800, and 990MFW and 13 authors at 900MFW. I used CoV Tuning on this word list using whole texts, retained words found in 7 or more texts, and then tested the 4,000-word sections. The improvement over the standard method is less dramatic here: the 600MVW succeeded for just 10 authors, the 700 for 12 authors, and the

800, 900, and 990 for 13. Minimums of 5, 6, or 8 texts are less effective.

Finally, I assembled a set of 40 late 19th and early 20th century novels by 20 authors, 2 novels each, intentionally selecting authors and texts known to be difficult to attribute. Standard cluster analysis correctly groups the texts by only 12 of the 20 authors, from 990 to 600MFW. I used CoV tuning on the word list as above, retaining only words found in at least 28 texts. Here, CoV Tuning matches the results for the standard method for the 700 and 800MVW, but correctly groups 13 authors for the 600MVW and 14 for the 900 and 990MVW. A minimum of 33 texts is less effective. Figures 5 and 6 show standard and CoV Tuned analyses.

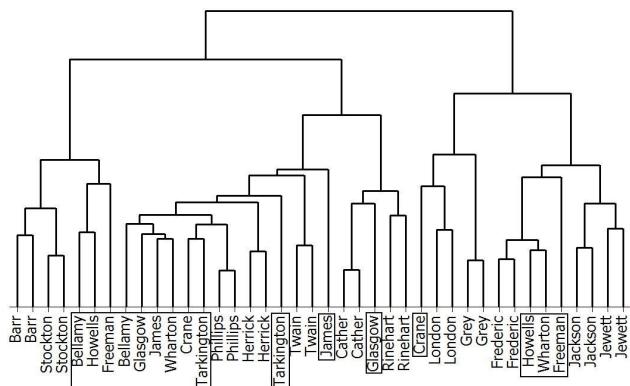


Fig. 5: Standard, 700MFW

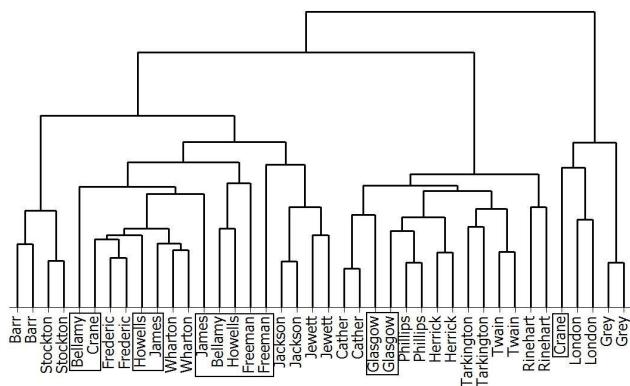


Fig. 6: CoV Tuned, 700MVW

More testing will be required to determine whether CoV Tuning gives consistently superior results on most sets of texts, and an automatic method for selecting the optimum limit to set on the minimum number of texts in each analysis is needed. However, CoV Tuning already seems to be a potentially valuable method of combining the information about word frequency that has long been the focus of authorship attribution and computational stylistics with the information about consistency of use on which recent methods like Zeta, Iota, and Full Spectrum analysis (Hoover 2013) are based.

References

- Burrows, J. F.** (1992). "Not Unless You Ask Nicely: The Interpretative Nexus Between Analysis and Information." *LLC* 7 (2): 91-109.**Burrows, J. F.** (2005). "Who Wrote *Shameless*? Verifying the Authorship of a Parodic Text." *LLC* 20 (4): 437-450.
- Hoover, D. L.** (2003). "Statistical Stylistics and Authorship Attribution: An Empirical Investigation." *Literary and Linguistic Computing* 16(4) 2001: 421-444.**Hoover, D. L.** (2003). "Multivariate Analysis and the Study of Style Variation." *LLC* 18(4), 2003: 341-60.
- Hoover, D. L.** (2004). "Testing Burrows's Delta." *LLC* 19 (4): 453-475.

Hoover, D. L. (2010). "Authorial Style," in Dan McIntyre and Beatrix Busse (eds), *Language and Style: Essays in Honour of Mick Short*, New York: Palgrave: 250-71.

Hoover, D. L. (2013). "The Full-Spectrum Text-Analysis Spreadsheet," in *Digital Humanities 2013*, Lincoln, NE: Center for Digital Research in the Humanities, University of Nebraska: 226-29.

Jockers, M. L., D. M. Witten, and C. S. Criddle. (2008). "Reassessing Authorship of the Book of Mormon Using Delta and Nearest Shrunken Centroid Classification," *LLC* 23 (4): 465-491.

McKenna, C. W. F., and A. Antonia. (1996). "'A few simple words' of Interior Monologue in Ulysses: Reconfiguring the Evidence." *LLC* 11 (2): 55-66.

Rybicki, J., and M. Eder. (2011). "Deeper Delta Across Genres and Languages: Do We Really Need the Most Frequent Words?" *LLC* 26: 315-21.

Rybicki, J., and M. Heydel. (2013). "The Stylistics and Stylometry of Collaborative Translation: Woolf's Night and Day in Polish," *LLC*, first published online May 27, 2013.

Making Waves: Algorithmic Criticism Revisited

Hoover, David L.

New York University, United States of America

In *Reading Machines: Toward an Algorithmic Criticism*, Stephen Ramsay suggests that computational literary studies remain marginalized because they lack "bold statements, strong readings, and broad generalizations" (2011: 2). They are too cautious, too scientific, to interest literary critics, who value opening texts to new interpretations over solving problems (10-11). Ramsay suggests that a feminist discussion of *The Waves* challenges algorithmic criticism: "literary critical arguments of this sort do not stand in the same relationship to facts, claims, and evidence as the more empirical forms of inquiry. There is no experiment that can verify the idea that Woolf's . . . 'elision of corporeal materiality' exceeds the dominant Western subject" (7). Although many critical claims are computationally intractable, Woolf's "elision of corporeal materiality" surely has textual implications that might be tested computationally. Literary criticism's problematic relationship to facts, claims, and evidence seems more like a bug than a feature, but here I want to re-examine and interrogate Ramsay's algorithmic provocation.

Three male and three female characters in *The Waves* speak in alternating monologues, an experimental technique that has invited critical comment about what axes of difference or unity characterize the novel. "The 'problem' . . . with Woolf's novel is that despite evidence of a unified style, one suspects that we can read and interpret it using a set of underlying distinctions. We can uncover those distinctions by reading carefully. We can also uncover them using a computer" (Ramsay 2011: 10-11).

Ramsay treats the six monologues as a corpus of documents and investigates them with tf-idf, from the field of information retrieval: $tf^*(N/d)$. This, he suggests, should identify each monologue's characteristic words more effectively than a traditional word-frequency list. Tf-idf is the term's frequency (tf) multiplied by the total number of documents (N; here 6) divided by the number of documents containing the term (df; document frequency). Tf-idf reduces the importance of function words and increases the importance of speakers' characteristic words because the frequencies of words used by only one speaker are multiplied by six (6/1), while the frequencies of words used by all six speakers are multiplied by one (6/6) (Ramsay 2011: 11). After identifying each speaker's most characteristic words, he reveals that he has actually used the formula, $1 + tf * \log(N/d)$, which includes a log function (reducing the effect of a word's appearance in only one speaker), and adds 1 (preventing the measure from becoming negative). The purpose of the alterations "is not to bring the results into closer conformity with 'reality,' but merely to render the weighting numbers more

sensible to the analyst" (Ramsay 2011: 15). Yet the variants are not "merely" at the whim of the analyst; they have testable consequences.

But let us travel a bit further with Ramsay. He presents the words with the highest tf-idf scores in Louis's monologue (listed in Fig. 1, along with his tf-idf scores, my tf-idf scores, and their frequencies), and suggests that "Few readers of *The Waves* would fail to see some emergence of pattern in this list" (12). For example, *western* seems to echo Louis's concern about his *Australian accent*, and *England* (all top 25). But actually *western*, *wilt*, and *thou* appear only in Louis's quotations from a sixteenth-century poem. Ramsay's provocation ignores some interesting questions: should Louis's quotations be considered his speech (and retained?), or the anonymous author's (and omitted?).

Freq.	Ramsay tf-idf	Word	My tf-idf
10	5.917438	mr	5.917438
9	5.728657	western	5.728657
8	5.517619	nile	5.517619
6	5.002161	australian	5.002161
5	5.002161	beast	4.675485
6	5.002161	grained	5.002161
6	5.002161	thou	5.002161
6	5.002161	wilt	5.002161
5	4.675485	pitchers	4.675485
5	4.675485	steel	4.675485
4	4.275666	attempt	4.275666
4	4.275666	average	4.275666
4	4.275666	clerks	4.275666
4	4.275666	disorder	4.275666
14	3.916497	accent	3.997913
3	3.760208	beaten	3.760208
3	3.760208	bobbing	3.760208
3	3.760208	custard	3.760208
3	3.760208	discord	3.760208
3	3.760208	eating-shop	3.760208
3	3.760208	england	3.760208
3	3.760208	eyres	3.760208
3	3.760208	four-thirty	3.760208
3	3.760208	ham	3.760208
3	3.760208	lesson	3.760208

Fig. 1: Louis's Most Characteristic Words

Trying to recreate Ramsay's analysis reveals further interesting points. The mismatched tf-idf scores (bold) reflect different word frequencies. His score for *beast* requires 6 occurrences, not 5, but the 6th is in the omniscient narration. His score for *accent* requires 13 occurrences, not 14, but Rhoda's most characteristic words include *them-* and *accent* occurs once as *accent-*, which presumably reduces his count by 1. (What constitutes a word is a surprisingly complex question, but treating *accent-* as a word seems odd.) The rarity of the words shows how strongly tf-idf privileges words limited to one character (only *accent* appears in 2). Ramsay's intervention raises interesting questions: What does it mean to choose this algorithm? How do the results affect our emerging reading of *The Waves*? (Ramsay 15). But how to answer these questions?

Even the identification of characteristic words is problematic. *Tick* and *hoot* both occur 4 times, only in Bernard, but all 8 occurrences are in 3 consecutive sentences. How

"characteristic" is that? Low occurs just 5 times, "only" in Bernard, yet it also occurs in the omniscient narration. Analyzing only the 6 characters seems reasonable, but should Bernard's characteristic words also occur in the narration? (Some consider Bernard to be modeled on Woolf herself [Ramsay 2011: 13].) Including the narration seems both intriguing and problematic, not least because it is not dialogue. Doing so removes *low*, *canopy*, *bowed*, and *brushed* from Bernard's most characteristic words and *beast*, *steel*, and *discord* from Louis's. What questions does this raise?

Most algorithms for computational approaches come from authorship attribution, where ostensibly correct answers exist. But Ramsay is right that the existence of "correct" answers to questions like "Do the men and women speak differently?" or "Do the characters have distinct and consistent voices?" is precisely at issue. Examining *The Waves* in the light of Ramsay's provocation raises so many intriguing questions that they cannot all be addressed here. But we can approach the question of character individualization by using a radical deformation. I randomly sorted the lines of the six monologues, then selected the first 6067 words of each, the length of the shortest monologue (Susan's). I identified the 50 most characteristic words using Ramsay's tf-idf formula, and tested how well they group with the remainders of the longer monologues using cluster-analysis, starting with all 300 words (in descending tf-idf order), then reducing the number gradually. The best result, for the 20 most characteristic words, is shown in Fig. 2.

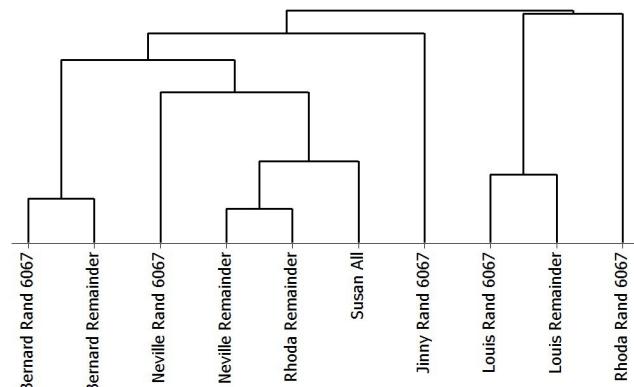


Fig. 2: Tf-idf and Character-Individualization

Bernard's and Louis's sections group together, while Neville's and Rhoda's fail (Jinny and Susan have too little text for 2 sections). A simple word frequency list, however, correctly groups all 4 in many analyses (see Fig. 3, based on the 300 mfw), providing a tentative answer to the question of distinct voices.

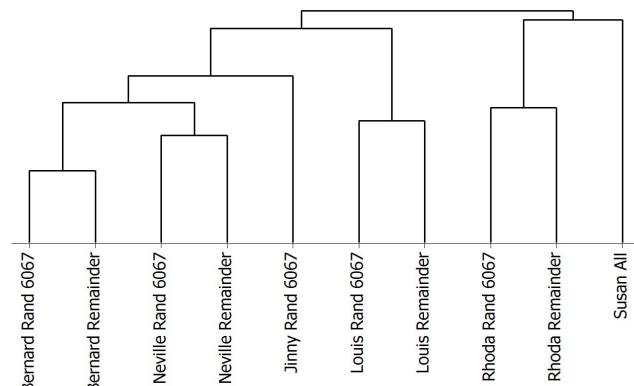


Fig. 3: The Most Frequent Words and Character-Individualization

Selecting the 50 most characteristic words for each monologue using Zeta (Craig and Kinney 2011), also produces many perfect results (see Fig. 4, based on the 200 most characteristic words). This very different method, which measures consistency of use rather than frequency, confirms

the distinctness of the voices. Finally, testing the six characters in 2,000-word sections with 2-grams (based on the six full monologues) also yields many completely correct clusters (see Fig. 5 for an analysis based on the 900 mf2Grams).

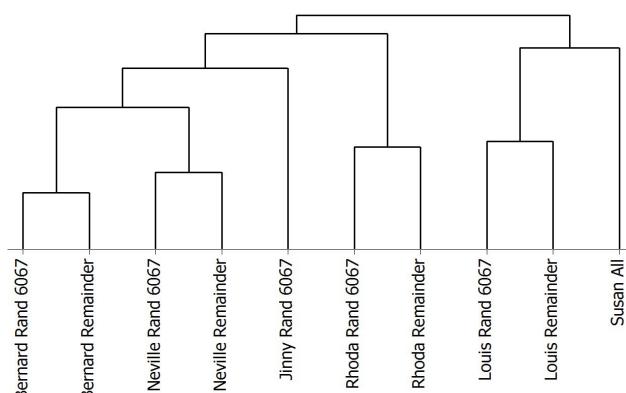


Fig. 4: Zeta and Character-Individualization

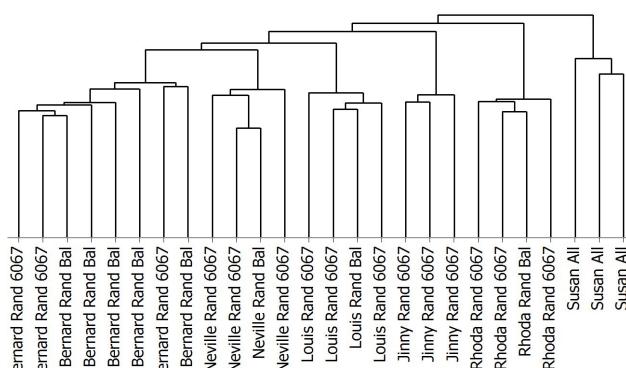


Fig. 5: 2Grams and Character-Individualization

Ramsay suggests that treating the question of whether the six characters in *The Waves* share “the same stylistic voice” as a problem to solve is a “category error,” and that the proper question—one computers cannot answer—is “Can I interpret (or read) it this way?” (2011: 9–10). Critics still can read the novel as a single stylistic voice, and the six monologues undoubtedly share many characteristics, but, in spite of a host of very interesting remaining questions about the status of algorithms, arguments, and evidence, it seems reasonable to make the bold claim that there are six distinct character voices in *The Waves*.

Ramsay’s provocative intervention is valuable for forcing us to re-examine our methods and focus on questions of interest to traditional literary scholars. But further analysis of his provocation and his algorithm suggests that more attention to the text, to the nature and function of the algorithms, and to method may prompt bold claims that rest on a sounder foundation. Further work will help us explore the boundary between computationally-tractable and computationally-intractable questions and the significance of that boundary.

References

Ramsay does not analyze, then interpret, a method criticized by Fish (2012). He begins with a literary judgment, then investigates it. Despite Fish’s criticism, both methods seem valuable.

Ramsay actually uses $(1+\ln(\text{tf})) * \ln(N/\text{df})$, not $1+\text{tf} * \log(N/\text{df})$. Many tf-idf formulas exist; for the 4 I tested, 20 of the 25 most characteristic words are the same.

It also reveals that Ramsay has (not unreasonably) omitted Bernard’s final long “summing up” chapter, as I have also done.

The characters do not group completely by gender in the graphs above, also suggesting a tentative answer to “Do the men and women speak differently?” that is different from Ramsay’s.

Craig, H., and A. Kinney. (2010). *Shakespeare, Computers, and the Mystery of Authorship*. Cambridge UP.

Hoover, D. L. (2007). “The End of the Irrelevant Text: Electronic Texts, Linguistics, and Literary Theory,” *Digital Humanities Quarterly* 1(2).

Ramsay, S. (2011). *Reading Machines: Toward an Algorithmic Criticism*. Urbana: University of Illinois Press.

Fish, S. (2012) “Mind Your P’s and B’s: The Digital Humanities and Interpretation,” *The New York Times*, online, January 23..

The Workspace for Collaborative Editing

Houghton, Hugh

h.a.g.houghton@bham.ac.uk

Institute for Textual Scholarship and Electronic Editing (ITSEE), University of Birmingham

Sievers, Martin

sievers@uni-trier.de

Trier Center for Digital Humanities (Kompetenzzentrum für elektronische Erschließungs- und Publikationsverfahren in den Geisteswissenschaften, Universität Trier)

Smith, Catherine

c.j.smith@bham.ac.uk

Institute for Textual Scholarship and Electronic Editing (ITSEE), University of Birmingham

The Workspace for Collaborative Editing is a project funded by the AHRC (UK) and DFG (Germany) between September 2010 and December 2013. It has the goal of creating an online workspace to support the production of the *Editio Critica Maior* of the Greek New Testament by teams based in Birmingham, Münster and Wuppertal and collaborators dispersed all over the world.¹ The edition has been in progress since the 1990s, but the obsolescence of key tools and encodings have led to this ambitious project to connect all the different stages of the editorial process through online interfaces and shared databases.

The production of a critical edition involves the identification and selection of manuscripts to be included, the acquisition of images, the creation of full-text electronic transcriptions (which are themselves published as separate electronic editions, linked to the electronic apparatus, enabling further research in related fields), the automatic comparison of these transcriptions to generate a critical apparatus of all variant readings, the editing of this apparatus by scholarly editors to filter out ‘noise’ and prepare the data for analysis using genealogical tools, the addition of evidence from early translations and biblical quotations and the publication of the material in electronic and printed form.

The aim of the Workspace project has been to adopt existing standards and open-source solutions in order to create a lightweight architecture capable of being easily renewed and updated, so that both the data and software created may be reused by other projects. The result is an open-source browser-based environment written in Python and Javascript. The core software consists of a MongoDB database and the asynchronous web application framework MAGPY. Data is stored in JSON and made available via a RESTful interface. On top of this is a layer of applications which call the relevant data objects for the individual editing processes. The goal of transparency at every level of the editing process means that a record is kept of each object at each stage of the process, and any modifications introduced are treated as additional records rather than replacing existing data.²

The Greek New Testament provides a very specific use-case, with a large amount of data already created and highly developed editorial principles. In addition, ongoing work by existing editorial teams offers the opportunity for immediate testing in real-life situations. Developing in these circumstances can be a challenge, with the evolution of guidelines, changes of editorial practice and ‘creeping featurism’. The system needed

to make existing legacy data compatible with the much more detailed XML encoding developed by the project and cater for as many known and potential scenarios as practicable. The dispersed team of editors was often called upon to codify their procedures and reach a common mind on problems presented by live data, including agreeing changes in policy. As a result, the creation of the Workspace has proceeded hand in hand with the development of different stages of the edition as a whole.

The two principal areas in which the Workspace meets a pressing need are the development of a transcription editor, which produces and allows the editing of valid XML in a WYSIWYG environment, and a collation editor which enables the scholarly creation of a critical apparatus. Both of these are browser-based, in order to enable dispersed collaborators to work with differing operating systems and contribute directly to the central data store.

The Transcription Editor has been created by team members at the Trier Center for Digital Humanities and released as open-source at the end of the project.³ Its basis is the platform independent TinyMCE package.⁴ A set of options for mark-up was then developed through a series of menus and shortcuts (cf. Figure 1). The aim is to allow student and volunteer transcribers not familiar with XML to work in an environment which matches as closely as possible the format of the transcriptions already published in the system. The mark-up in the browser uses HTML encoding. An export function converts this into XML matching the specifications developed by the project.⁵ Likewise, an import function is required in order to support the editing of existing transcriptions. Some of the problems include the encoding and display of paratextual information, normally located in the margins of a manuscript. The dialogue box for entering this information has to have the same functionality as the main transcription interface for recording unclear or supplied text, corrections and so on. The concept has therefore been developed of the “editor-within-an-editor” which makes this possible. A problem with the import of existing transcriptions is the sequence within which elements were nested within the XML. As a result, it has been necessary to establish a system of tag sequences supported by the editor. The standalone nature of the Transcription Editor and its use of an agreed set of TEI encoding means that it can be installed as a plug-in to different environments, including the New Testament Virtual Manuscript Room (NTVMR 2.0)⁶ as well as the Workspace for the production of the critical edition.



Fig. 1: The Transcription Editor in the NTVMR environment. Based on the selection different menu options for breaks, corrections, deficiency, ornamentation, abbreviations, marginalia, notes and punctuation are offered. Mouseovers and different colours help the users to identify different structures.

The Collation Editor provides an interface to the CollateX engine developed by the INTEREDITION project, the successor to the COLLATE program by Peter Robinson.⁷ This software performs one of the most mechanical and error-prone tasks in an edition, namely the comparison of all witnesses in each variation unit to build up a critical apparatus. Each file is aligned using an algorithm taking into account not just spelling variations, additions, omissions and substitutions, but also transpositions within each block of text. However, the output still requires considerable input from scholars in order to clean up the raw data for publication as a critical apparatus. The first stage is regularisation, the elimination of insignificant variations such as spelling errors. The variant readings are set out underneath a base text, with the witnesses attesting each

reading visible in a mouseover box (see Figure 2). An interface built using the redips drag-and-drop library allows editors to drag-and-drop the readings for regularisation onto the correct form.⁸ For each regularisation, a dialogue box requires users to state the scope of the regularisation and also its nature. Once this is completed, the regularisation is marked in grey and a rule is saved to the database. The ‘recollate’ button sends the data back through CollateX, preferring the regularised token to the original form where present. This means that a different configuration of readings may appear in each column, as the data is cleaned up and a better match is made by the collation algorithm. The second stage involves setting the length of each variant unit, again implemented through a user-friendly drag-and-drop interface for combining or splitting neighbouring columns. One of the dangers with this interface is changing the overall sequence of words in a manuscript by combining different units and repositioning readings. A checking mechanism has therefore been developed which warns the user as soon as the sequence of any manuscript has been disrupted. On some occasions, the data is best displayed as two units of different lengths. By right-clicking on the relevant reading, it can be sent to a line below as an “overlapping variant”, which can then be combined and manipulated like the other columns. One further complication is that an overlapping variant such as a lengthy transposition of words may also contain a reading which should cited in the main sequence. The system therefore makes it possible to duplicate such readings. The final stage is the ordering of variant readings within each unit and assigning the appropriate reading identifier. From here, the apparatus can be output in a number of forms, such as a positive or negative plain text apparatus, an XML encoded apparatus, or a set of values for incorporation into a database for phylogenetic analysis. The information added in the regularisation dialogue box makes it possible to generate automatically the lists of original forms for orthographic variants and erroneous readings which are printed in an Appendix in the *Editio Critica Maior*.

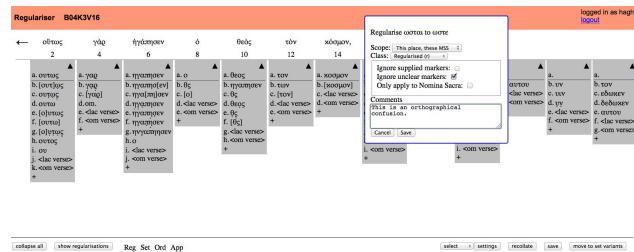


Fig. 2: The regularisation interface with the dialogue box displayed.

The presentation will briefly demonstrate the Workspace, especially the two interfaces described above. We will discuss some of the problems encountered during its development, along with their solutions. Although the scope of the original project was specifically to support an edition of the Greek New Testament, a pilot project to customise the environment for an edition of Avestan texts will be outlined: from here, we hope that it will be possible to develop the Workspace for use with other textual traditions.

References

For the history of the ECM project, see **Klaus Wachtel**, *Editing the Greek New Testament on the Threshold of the Twenty-first Century* “Literary and Linguistic Computing 15 (2000), 43–50; **D.C. Parker and Klaus Wachtel**, *The Joint IGNTP/INTF Editio Critica Maior of the Gospel of John: its goals and their significance for New Testament scholarship*, presented at the Annual Meeting of SNTS, August 2–6, 2005, Halle. epapers.bham.ac.uk/754/.

2. For documentation, see zeth.github.io/magpy/index.html. The source code may be downloaded from github.com/zeth/magpy.

3. sourceforge.net/projects/wfce-ote/

4. www.tinymce.com/
5. **H.A.G. Houghton**, *The Electronic Scriptorium: Markup for New Testament Manuscripts*, in Claire Clivaz, Andrew Gregory and David Hamidovic (edd.), *Digital Humanities in Biblical, Early Jewish and Early Christian Studies*, Leiden: Brill (2013), pp. 31–60; the latest version of the specifications is at epapers.bham.ac.uk/1727.
6. ntvmr.uni-muenster.de/en_GB/transcribing
7. **P.M.W. Robinson** (1994), *Collate: Interactive Collation of Large Textual Traditions, Version 2. Computer Program distributed by the Oxford University Centre for Humanities Computing*, Oxford. The history of Collate is described by Robinson in a 2007 blog post at: www.sd-editions.com/blog/?p=15. For CollateX, see www.interedition.eu/. The source code is available from collatex.net/.
8. www.redips.net/javascript/drag-and-drop-table-content/

Enjambment and the Poetic Line: Towards a Computational Poetics

Houston, Natalie
nmhouston@gmail.com
 U of Houston

The large-scale digitization of public domain texts carried out in recent years by Google and university libraries offers broader scope for literary-historical research into the development and cultural function of specific literary forms and genres. Traditional scholarship on select canonical texts can now be combined with the computational analysis of large document sets to provide insight into what makes those texts distinctive or representative of larger historical patterns. This paper discusses work in progress from *Understanding Victorian Poetic Style*, a project that examines how text analysis methods can be adapted to the study of poetics at the large scale. Stylistic analysis has largely been focused on identifying the distinctive linguistic patterns used by particular authors; I'm interested in extending these methods to examine poetic genre and form as shared historical, cultural practices. To do this requires attending to the multiple ways that poems create meaning through deliberate structures, such as enjambment, repetition, and rhyme.

Recent scholarship in nineteenth-century poetry has returned to the cultural study of form and to the study of historical poetics, which examines the history of theories about poetry's linguistic textures.^{1 2 3} The nineteenth century produced a tremendous variety of metrical, rhymed, and stanzaic verse as well as free verse, which does not follow a set meter or rhyme pattern. The digitization of nineteenth-century texts now affords the possibility of contributing to our historical understanding of poetic form with large scale analyses of poetic practice. This paper presents my current research into using computational text analysis for understanding the historical practice of enjambment, a key feature of the poetic line. This research contributes both to the project of sociological poetics, the understanding of literary form in its broadest historical function within human culture, and to the development of a computational poetics.⁴

The poetic line is a distinguishing feature that separates poetry from prose. Like prose, poems contain sentences which can be analyzed syntactically and semantically. But in verse those sentences are arranged in lines. The poetic line is defined rhythmically in metrical verse; is marked through sound in rhymed verse; and is visually reinforced by white space in free verse and indeed in printed poems of all kinds.

Lines of poetry are defined to a large degree by their endings: some lines are firmly "end-stopped" by closing punctuation, like a period, and some "enjambed" lines deliberately continue into the line which follows, often by breaking in the middle of a syntactic clause. Rather than seeing these as simple oppositions, John Hollander looks at the common notation for marking poetic line endings when quoted in the midst of prose (i.e., Wordsworth's "I wandered lonely as

a cloud / That floats on high o'er vales and hills,") and proposes that we understand enjambment as:

" . . . a kind of spectrum, along which we would arrange all the possible ways of terminating lines, considered not as boundaries or termini, but as the kinds of cutting into syntax which the slant-dash notation illustrates" (99).⁵ Enjambment can thus be understood as one measure of the relation between the poetic line and the syntactic sentence. As T.V.F. Brogan suggests, "The sense in prose flows continuously, while in verse it is segmented so as to increase information density and perceived structure."⁶ Poetic style can only be partially understood through "bag of words" or even n-gram analyses, since those words are deliberately arranged not only in sentences but in lines. Developing quantitative measures for this segmentation contributes to an historical poetics that defines form and genre as cultural phenomena.

This paper describes my current approach to computationally analyzing enjambment in a corpus of poetry published in England between 1840-1900. It compares the utility of three different measures of the relation between the poetic line and the syntactic sentence:

- (1) a simple line:sentence ratio;
- (2) a spectrum definition marking degrees of enjambment based on different kinds of punctuation;
- (3) a spectrum definition marking degrees of enjambment based both on different kinds of punctuation and part-of-speech tagging.

These measures are considered as features of poetic style that can be used alongside other features in classification experiments or as markers of historical change in poetic practice. This computational analysis can contribute to our understanding of enjambment as a feature of an individual poets' style; as a feature of particular poetic forms, themes, and genres; and as a feature of poetic discourse in particular historical periods. As James Scully suggests:

"There is no unpositioned, dehistoricized technique – no way to comprehend line breaks in and of themselves. It's not simply that these are socially specific practices, nor that there are different kinds of line breaks . . . but that line breaks do not work the same way in ballad quatrains as in blank verse, nor in prescribed verse as in free verse. . . Free verse too has a lineage, a historically reproduced repertory of conventions through which it works and to which it responds" (108).⁷ The computational analysis of enjambment as it occurs in both prescribed and free verse in the nineteenth century can help us better discover and understand that repertory of formal conventions. By moving the study of poetics to the large historical scale, computational analysis can begin to generate a more detailed historical account of the historical and cultural functions of poetic form.

References

1. Levinson, M. (2007). *What is New Formalism?*, PMLA: Publications of the Modern Language Association 122.2: 558-69
2. Hall, J. D., ed. (2011). *Meter Matters: Verse Cultures of the Long Nineteenth Century*. Athens: Ohio University Press.
3. Hollander, J. (1975). *Vision and Resonance: Two Senses of Poetic Form*. New York: Oxford University Press.
4. Bakhtin, M. M. and Medvedev, P. N. (1978). *The Formal Method in Literary Scholarship: A Critical Introduction to Sociological Poetics*. Wehrle, A. (trans). Baltimore and London: The Johns Hopkins University Press.
5. Hollander, J. (1975). *Vision and Resonance: Two Senses of Poetic Form*. New York: Oxford University Press.
6. Scully, J. (1988). *Line Break*. In Frank, R. and Sayre, H., (eds), *The Line in Postmodern Poetry*. Urbana and Chicago: University of Illinois Press, pp. 97-131.
7. Scully, J. (1988). *Line Break*. In Frank, R. and Sayre, H., (eds), *The Line in Postmodern Poetry*. Urbana and Chicago: University of Illinois Press, pp. 97-131.

A glimpse of the change of worldview between 7th and 10th century China through two leishu

Hsiang, Jieh

jhsiang@ntu.edu.tw
National Taiwan University

Chen, lihua

National Taiwan University

Chung, Chia-Hsuan

National Taiwan University

What is *leishu*

Leishu (類書) is a unique form of Chinese source book for quick references and quotations. The editor of a *leishu* would collect a large number of books, develop a knowledge structure of the intended knowledge domain (usually with a number of categories that cover the domain, each with a list of subjects), then extract texts, that the editor deemed relevant to each subject, from the content in the books. Thus each subject has a list of *entries* which are texts taken from existing books. Because the purpose was for quotation, the meanings of the subjects were not explained and the texts were quoted verbatim, usually with the source indicated. (The lack of explanation of the subjects differentiates a *leishu* from an encyclopedia.) The earliest *leishu*, *Huanglan* (皇覽), dates back to 220AD.

During the past 300 years *leishu* had been looked down upon by many Chinese scholars. Its compilation nature was regarded as a lack of originality. The ease of use as quick reference was criticized as providing a short cut and thus encouraged shallowness in scholarship. Its value in the scholarly circle had largely been limited to being a vessel from which fragments of lost books were collected.

The advance of information technology allows us to look at *leishu* from a very different angle. The categories and subjects, together with the selections of entries, reflect how the world was perceived at the time when the *leishu* was compiled. Comparing two *leishu* from different era, thus, provides a way to observe how the world view had changed during the years between the two tomes. This was not possible until now, when the availability of *leishu* in searchable full-text form finally provides a way to study and compare the books in their entirety.

In this paper we present such a study. The *leishu* we have chosen are *yiwenleiju* – YL (藝文類聚), completed in 624AD (early Tang dynasty) by Ouyang Xun, and *taipingyulan* – TY (太平御覽), completed in 984AD (early Song dynasty) by Li Fang.

Both were commissioned by the emperors and were about the general knowledge of the world. YL divides its view of the world into 46 categories, further into 734 subjects. TY has 55 categories and 5,597 subjects.

It was stated in the preface of TY that YL, written 350 years earlier, was among the three main *leishu* consulted. (The two other no longer exist.) With the full texts available through digitization, we can finally check effectively the inheritance relation between TY and YL. It also provide a way to observe the change of the world view between the two books, which is reflected not only by the numbers and titles of the categories, but also by the subjects that each category covers, and the entries that are listed under the subjects. New subjects indicate the emerging importance of new concepts; the increase of entries under the same subject means more knowledge (or interest) about the subject; and the assignment of the same entry to a different subject signals the change of a viewpoint. We will give some preliminary findings in this paper.

Processing the texts and building a system for comparison

The full texts in our study are obtained through Guoxuewang (www.guoxue.com). Treating an entry as a basic unit, we parsed each entry into the content and its source (the book from which the text is taken). The source is further analyzed to identify the title of the book (sometimes with the chapter and the section), the author (if known), and the era (dynasty). An XML format is designed to associate an entry with its category, subject, source, and content. Further analysis was conducted to resolve name and author conflicts. According to our results, YL has 14,572 entries, extracted from 5,628 sources, of which 787 are books and the rest (4,841) are individual articles such as poems, letters, and memorials. TY, on the other hand, has 65,633 entries, extracted from 2,327 books and 1,832 other sources. Among the books 629 are cited by both. Of the remaining 1,698 books that are cited by TY but not YL, 498 are from pre-Tang era (which could have been included in YL) and 980 from era unknown (although the majority should be pre-Tang). The number of books that are certain to have been written after YL was completed is only 220. The total numbers of words in the two books are about 900,000 and 4,000,000, respectively.

We then designed an algorithm based on the longest common sequence method to check if an entry appears in both books. Modifications to LCS are necessary to deal with different styles or errors in the quoted texts. Since the entries of YL are classified into *affairs* (事) and *literature* (文) while TY is primarily about *affairs*, we focused initially on comparing the entries in YL about *affairs*. We found that among the 9,701 such entries in YL, 7,249, or 75%, appear in TY. That means TY did indeed heavily reference the earlier book. Those entries appeared 11,022 times in TY, since an entry may appear more than once.

We built a visualization system to compare the two books. An important feature is to show how many entries of a category/subject of one book are also cited in the other book and how they are distributed. We present here an example using the category *fuming* (divine signs, 符命) in YL. Figure 1 shows that of the 41 entries of the affairs part of *fuming* in YL, 27 also appeared in TY (left-most column). The middle-left column shows that these 27 entries appeared 51 times in TY. The third column indicates that these 51 appearances belong to 40 subjects, and the right-most column further indicates that they belong to 16 categories.

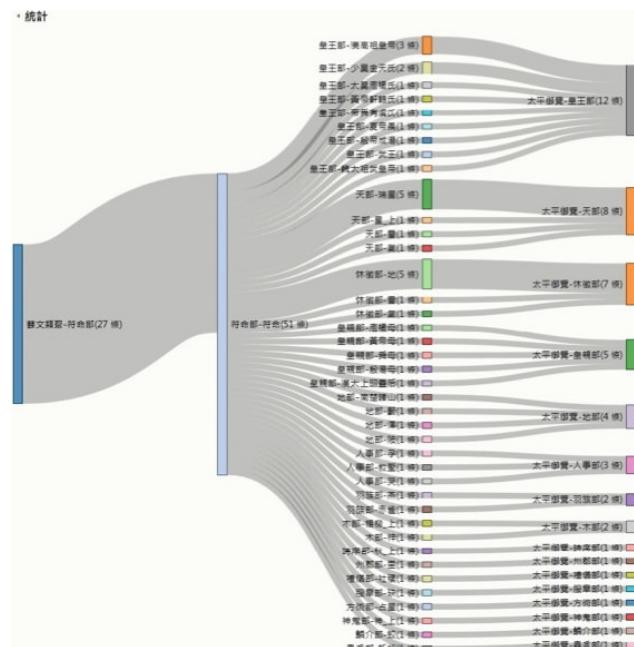


Fig. 1: Correspondence of entries of fuming in YL and TY

Figure 2 is the list of categories with the number of entries that match those in *fuming* of YL. Each category can be expanded to show the subjects that contain some of the related entries.

對應到太平御覽各部目的對應狀況

- 太平御覽-皇王部：12個對應
- 太平御覽-天部：8個對應
- 太平御覽-休徵部：7個對應

休徵部-地：	5對
休徵部-氣：	1對
休徵部-雷：	1對

- 太平御覽-皇親部：5個對應
- 太平御覽-地部：4個對應
- 太平御覽-人事部：3個對應
- 太平御覽-木部：2個對應
- 太平御覽-羽族部：2個對應
- 太平御覽-蟲豸部：1個對應
- 太平御覽-服章部：1個對應
- 太平御覽-方術部：1個對應
- 太平御覽-時序部：1個對應
- 太平御覽-禮儀部：1個對應
- 太平御覽-州郡部：1個對應
- 太平御覽-鱗介部：1個對應
- 太平御覽-神鬼部：1個對應

Fig. 2: Categories and subjects of TY that contain entries of fuming from YL

Figure 3 lists the entries of fuming in YL and the corresponding (similar) entries in TY.

【藝文類聚-符命部-符命】
1082.【藝文類聚】符命部-符命《春秋漢澤巴》： ◎顯示相似條目 里社場，此異冉姓人，其名則百姓歸之。社，異之貴也。堪則教令行，堪聖人能之。均，堪之怒也。
9099.【太平御覽】州郡部-里《春秋漢澤巴》 里社場，此異冉姓人，百姓歸之。
25731.【太平御覽】人部-叔靈《春秋漢澤巴》 里社場，此異冉姓人，百姓歸之。
32941.【太平御覽】禮樂部-社靈《春秋漢澤巴》 里社場，此異冉姓人也，百姓歸之。
54792.【太平御覽】休徵部-地《春秋漢澤巴》 里社場，此異冉姓人，其別，百姓歸之。
1883.【藝文類聚】符命部-符命《春秋合圖》： 禹母冉節，出觀三河，忽然強至，赤龍負禹新生，生禹。 1884.【藝文類聚】符命部-符命《河圖》： ◎顯示相似條目 禹母扶都，見白禹貢月，慈感而生禹。
7319.【太平御覽】皇帝部-殷帝成祖《河圖》： 扶都見白禹貢月，或言黑帝帝。
8107.【太平御覽】皇親部-符漢母《河圖書命》 扶都見白禹貢月，慈感，生禹帝子灌。
1885.【藝文類聚】符命部-符命《春秋元命苞》： ◎顯示相似條目 殷討之時，五蠻朝於禹，禹省諸神之精，周據禹與，禹起於禹，而五蠻朝之，得天下之祚。
568.【太平御覽】天部-星上《春秋元命苞》 禹討之時，五蠻朝於禹，禹者，諸神之精，周據禹與。
1886.【藝文類聚】符命部-符命《尚書中統》： ◎顯示相似條目 季秋，赤帝征丹青，火星，止於禹戶，禹拜稽首，禹應聲，禹嘆曰：近禹當帝子。
2520.【太平御覽】序部-秋上《尚書中統》 禹主生為西伯，季秋之月甲子，赤帝征丹青入夏庭，止於禹戶，乃拜稽首受敕，曰：“近禹當帝子，亡振者，封也。”

Fig. 3: Comparisons of entries

Evolution of worldview

Comparing the two *leishu* reveals a number of striking changes in worldview between 7th and 10th century China. We briefly describe three of them.

(1) The disappearance of fuming – divine justification

Fuming, a sign that provides divine justification to overthrow a dynasty (and to establish a new one), was an extremely important concept during Han Dynasty (206BC – 220AD).

Indeed, *Fuming* was included in YL as a category. However, although the number of categories increased from 46 to 55, *fuming* conspicuously disappeared from TY (it did not even appear as a subject). Although many of the entries are still included in the latter book, they are listed in subjects such as an emperor or his mother and are scattered over 16 categories. This means that the very concept of *fuming* became irrelevant politically at early Song. Other subjects such as auspicious signs (祥瑞) met similar fate. Song Confucianism went through a dramatic development in mid 11th century. The disappearance of *fuming* from TY seems to hint that the tide of ridding mysticism may have started a lot earlier, since TY was completed in 984AD.

(2) The rise and fall of Daoism and Buddhism

In YL, entries about Buddhism were listed in the category *inner canon* (內典), a Buddhist-centric term; non-Buddhist religious books are outer canons. Entries about Daoism appeared under the subject *immortals* (仙道) in the category *supernatural* (靈異). In TY, Buddhism and Daoism are both categories, with 10 and 53 subjects each. While the number of entries about Buddhism increased from 169 to 197, those about Daoism went from 137 to 1402. The 53 subjects include detailed classification of garments and buildings for different religious purposes.

Buddhism was introduced to China during the 1st century, and became the national religion during the 5th century. Buddhism was the mainstream religion in early Tang, when YL was compiled. After mid Tang, the indigenous Daoism started a revival and became an organized religion. This trend, coupled with the anti-Buddhist attitude of the royal house and many intellectuals, weakened the dominance of Buddhism in China. This phenomenon is vividly demonstrated in our comparison.

(3) The emergence of foreign peoples

Although the concept of *foreign peoples* (四夷) in China is as old as 1000BC, there was no category nor subject about any foreign peoples in YL. For instance, Wuo (倭), a country probably located in modern day Kyushu Japan, started paying tribute to China as early as late Han dynasty (around 100AD). However, Wuo was not mentioned as a subject in YL, although appeared (3 times) in subjects describing produce such as mulberry. TY, on the other hand, has under the category foreign peoples 388 different countries and peoples with 920 entries, which include countries as far as the Roman Empire (大秦) and as near as various indigenous (yet alien) ethnic groups within China proper. The entries are from books spanning across 1,500 years and are mostly about foreign tributes, trades, and mythologies. Tang was a melting pot and its capital, Changan, was a great cosmopolitan city. The intimate and frequent interactions with foreign people apparently also contributed to the establishment of China as an individual and separate identity.

Discussion

The intellectual progress of China between 7th and 10th century had been described as “mediocrity in an era of prosperity”. In this paper we showed that the Song Confucianism revival of the 11th century might have had its way paved during this period. Our observations are mainly through comparing two *leishu*, *yiwenleiju* completed in 624AD and *taipingyulan* in 984AD. Such a study, however, would not have been possible without a systematic way to compare the knowledge structures and the contents of the two tomes. This paper presents such a method to tackle this prohibitive task.

References

- Thomas H. Cormen, Charles E. Leiserson, Ronald L (2001). Rivest and Clifford Stein. Introduction to Algorithms (2nd ed.). MIT Press and McGraw-Hill. pp. 350–355. ISBN 0-262-53196-8.

- Chung, Chia-Hsuan** (2013), A study On the Evolution of Classifications in Leishu Yiwenleiju and Taipingyulan through Textual Analysis, MS Thesis, National Taiwan University.
- Fu, Daiwie** (2007), "The Flourishing of *biji* or Pen-Notes Texts and its Relations to History of Knowledge in Song China (960-1279), in Qu'etait-ce qu'écrire une encyclopédie en Chine, pp 103-130, Presses Universitaires de Vincennes, Saint-Denis.
- Ge, Zhaoguang** (1998), Chinese World of Knowledge, Thoughts, and Believes Before 7th Century, Fudan University Press, Shanghai.
- Gernet, Jacques** (1996), A History of Chinese Civilization, Cambridge University Press.
- Hu, Daojing** (1981), Zhongguo gudai leishu, Zhonghua Books.
- Nakatsuhama, Wataru** (1974), *Yiwenleiju Yinshu Suoyin*, Wenguang Publishing, Taipei.
- Nie, Chongqi** (1934), *Taipingyulan Yinde*, publisher unkown.
- Wen, I-duo** (1956), *Tangshi zalun*, Shanghai Guji Publisher.
- Zhou, Shengjie** (1998), *Taipingyulai yanjiu*, Sichuan Bashu Books.
- Zurcher, Erich** (2007), The Buddhist Conquest of China, Leiensia Sinica, Vol 11.

Many subjects of TY contain sub-subjects. This figure counts all sub-subjects. If the sub-subjects are not included, then the number becomes 4,066.

As can be seen, both leishu utilized many sources that were not books. This is consistent with previous works, in particular the two important indexes: on YL by Nakatsuhama, which listed 1025 books and 4600 other sources, and on TY by Nie, which listed 5005 sources (without separating books from others). Both indexes contain mistakes, which explains why our numbers are different.

Li Fang, the main editor of TY, also compiled two other great leishu, *taipingguangji* (太平廣記) and *wenyuanyinghua* (文苑英華). Guangji consists of mainly stories and folklores and the latter literature. Since both tomes contain large quantities of material on both Buddhism and Daoism, one might get a different impression if comparing YL with these two leishu. However, as Fu pointed out, TY and guangji were derived from different *biji* (筆記) traditions and that TY "has an imperial center allied with literati" while guangji "lacks that political center and seems to have an other-worldly, spiritual orientation". Since TY "places a greater emphasis on dynastic orthodoxies", it is better suited for the comparison than the other two leishu.

Building impact and value into the development of digital resources in the humanities: Rhyfel Byd 1914-1918 a'r profiad Cymreig / Welsh experience of World War One 1914-1918

Hughes, Lorna
National Library of Wales, United Kingdom

Roberts, Owain
National Library of Wales, United Kingdom

McCann, Paul
National Library of Wales, United Kingdom

Introduction

In 2011, the National Library of Wales established a Research programme in Digital Collections. The research focus of the programme is to develop an understanding of use of our existing digital content, using this knowledge to identify ways that the content can be enhanced and made more valuable for use in research, teaching, or community engagement;

and building projects that develop new digital content that addresses specific research or education needs, in partnership with academics and other key stakeholders. This activity, an example of which is described below enables critical reflection on making digital objects and the impact of digital collections on the humanities. This research addresses all aspects of digital research methods in the arts and humanities, taking advantage of the convergent practices that are embedded in digital humanities to add impact and value to digital collections of Wales.

The impact of digital collections in the humanities

The NLW Research Programme has carried out research into the use of digital collections, using a variety of methods, especially those included in the TIDSR (Toolkit for the Impact of Digital Resources) developed by the Oxford Internet Institute (microsites.oiil.ox.ac.uk/tidsr/welcome). In 2011-12, we used these methods to carry out some analysis of the use of an NLW digital resource "Welsh Journals Online". The findings of this investigation were consistent with earlier investigations into the use of digital collections in the humanities. Significantly, the research carried out using this approach led to the conclusion that most studies attempt to measure the use of digital collections after they are launched.

From 2012-13, the Library led a JISC-funded collaborative initiative with other archives and special collections to digitise research material about the First World War in Wales. The result was Cymru 1914 (www.cymru1914.org), a freely accessible online resource containing 190,000 pages of archival materials (including photographs, manuscripts, artworks, and newspapers); 30 hours of audio and approx. 12 hours of audio-visual material. Approximately 30% of the content is in the medium of Welsh. This project has been an important opportunity to incorporate earlier findings about the factors that increase the use and impact of digital collections into each stage of the development of the project. This presentation will discuss this process, and present findings about the factors that increase the value of digital collections for scholarship, with recommendations about their implementation into the development of digital collections in the humanities.

Selection of content for digitization

The primary source materials for the Welsh experience of the First World War were fragmented, frequently inaccessible and difficult to access, yet collectively they form a unique resource of vital interest to researchers, students, and the public in Wales and beyond.

An extensive scoping process of the Library, Archive and Special Collections of Wales highlighted materials in Welsh collections with the greatest relevance to World War One that were suitable for digitization, based on demand for the analogue archive materials and bibliographic research to identify citations of key materials. Research themes were identified that crossed many disciplines, opening up new avenues of research and comparative history. The final refinement of the content selected was completed in consultation academics engaged in teaching and researching the First World War, assessing the content with the greatest value to future scholarship and incorporating considerations of IPR and copyright.

Ingest into an underlying technical repository

The project made content available through ingest into the NLW's Fedora-based digital repository architecture. This supports the archiving of multimedia digital content; and the further exposure process of content for harvesting and aggregation. Existing workflows were modified to allow ingest into the repository of new content types all content types created by the project: printed text; newspapers; photographs; manuscripts; audio, and moving image materials.

Interface development, including bi-lingual support

Users appreciate straightforward user interfaces, so a simple bilingual interface was developed to provide material in a variety of formats and at varying levels of archival complexity, while retaining the hierarchical structures of archives that increase usability – through familiarity – of digital resources.

An interface group conducted on-going, iterative usability testing and implementation, including several user workshops: a formative evaluation exercise; an education workshop; and a participatory design workshop, organized by the Humanities Research Institute at the University of Sheffield, who are working on a project entitled *Participating in Search Design: a study of George Thomason's English Newsbooks* (<http://>). The goal of the latter workshop was to see if the participatory design methodology could feed into development of the interface by engaging with potential end users[5]. This paper will supply data about the above activities and they will be presented with cross referral against specific user communities. As the resource is used information gathered will be used to generate user case studies.

Dissemination and stakeholder engagement

Stakeholder engagement throughout the development of the project was crucial to ensuring widest use and re-use of the content, via a process of collaboration and outreach to disparate user communities, and usability testing and engagement with the digital outputs of the project. The project team worked with core communities through an iterative process of engagement and input throughout the development of the project, through the establishment of a research network of academics using the content, specifically participating in the three stakeholder workshops, and five Community engagement workshops, organised by the People's Collection Wales (<http://www.peoplescollectionwales.co.uk>). Post-launch user data from products such as Google Analytics also shape our findings.

Sustainability

In many respects, the actions described above to promote use, uptake and embedding of the resource are the surest way to ensure sustainability: digital collections that are used will be sustained over the long-term as they become invaluable to education, research, and community building.[6] A recent report by ITHAKA for the Strategic Content Alliance, "Sustaining our Digital Future"[7] highlighted the need to make planning for sustainability a key component of the digital life cycle. The use of good practice in digitization, and the use of an open-source, scalable repository such as Fedora, is key to sustaining the digital objects, of course, but key to cultivating sustainability of our valuable digital content is to embed planning for impact into the planning and development of digital resources. Fedora is a vital component of our long-term sustainability plans, and our institutional setting is key to this. Providing a crucial resource for research, teaching and public engagement around the topic of the Welsh experience of the First World War will promote sustainability of the resource. A key factor in planning and designing the resource as described in this paper is to create a digital content platform that can be added to over time. We also plan to revisit the use of the resource and to use this summative evaluation as the basis for any required modifications to increase its use.

Conclusion

It is increasingly obvious that factoring in end use of digital resources as broadly as possible at the outset of a digitization project is crucial: impact is a crucial component of the entire digital life cycle. The ultimate use of digital materials is a consideration that impacts decisions made at every stage of this life cycle: selection, digitization, curation, preservation, and, most importantly, sustainability over the long term. The way that digital resources are used may be unanticipated at

the outset; or they may have value for different communities and disciplines than originally intended. The best resources have been developed in such a way that their use and re-use has been anticipated at the outset, and that unforeseen use is anticipated through the use of technical standards and approaches. Just as digital collections that have been developed in formats that are not "open" are far less likely to be re-used for teaching or research, if digitization is to have more impact than being a form of "digital photocopying", the user needs to be placed at the centre of the process from the outset.

References

- Hughes, L.M.** (2013) *"Digital Collections in the Humanities: Understanding Use, Value and Impact"*, special issue of 'Digital Studies / Le champ numérique' from SSHRC Montreal seminar on cyberinfrastructure
- Warwick, C., Terras, M., Huntington, P., Pappa, N., & Galina, I.** (2006). *The LAIRAH Project: Log Analysis of Digital Resources in the Arts and Humanities Final Report to the Arts and Humanities Research Council*). London: University College London. Available online: <http://www.ucl.ac.uk/infostudies/claire-warwick/publications/LAIRAHreport.pdf>.
- Hughes, L.M., Ell, P., Dobreva, M., and Knight, Gareth, K.** forthcoming (2013) *"Assessing and measuring impact of a digital collection in the humanities: An analysis of the SPHERE (Stormont Parliamentary Hansards: Embedded in Research and Education) Project"*, Literary and Linguistic Computing (Oxford)
- See Dobreva, M., O'Dwyer, A., and Konstantelos, L., in Hughes, L.M. (ed) (2011)***Digital Collections: Use, Value and Impact*. London: Facet
- For an overview of this approach, see: **Wessels, B., Dittrich, Y., Ekelin, A and Eriksen, S.** (2012). 'Creating synergies between participatory design of e-services and collaborative planning' in *International Journal of E-planning Research*, Volume 1, Issue 3, doi: 10.4018/ijepr.2012070101
- Hughes, L.M.,** (2011) *"ICT Methods for digital collections research*, chapter in Hughes, L.M. (ed) (2011) *Digital Collections: Use, Value and Impact*. London: Facet
<http://www.sr.ithaka.org/research-publications/sustaining-our-digital-future>

Using digitized newspaper archives to investigate identity formation in long-term public discourse

Huijstra, Hieke

h.m.huijstra@uu.nl
Utrecht University

Pieters, Toine

t.pieters@uu.nl
Utrecht University

This paper analyzes how digitized newspaper databases can be used in historical research on identity formation in public discourse. It discusses a new semantic text mining tool, Texcavator, which is currently being developed in the Dutch research program *Translantis: Digital Humanities Approaches to Reference Cultures*.¹ The paper presents a case study which combines the Texcavator tool with the publicly available Delpher² and with traditional historical methods in order to analyze identity formation of health risk groups in Dutch public discourse in the twentieth century. In particular, it focuses on the construction of the identity of people with excess body weight. Although the case is built around the Texcavator and Delpher mining tools and the newspaper database of the Dutch national library, the paper aims to investigate techniques to combine close and distant reading that can be transferred to other tools and repositories as well.

Newspapers are valuable sources in historical research. Until recently, however, investigating them was cumbersome

and time-intensive. The repositories of digitized newspapers now available in many countries solve many practical problems and offer wonderful opportunities, but they also introduce methodological problems of their own. (Bingham 2010; Nicholson 2013) Bob Nicholson has recently shown how digitalization enables us to approach newspapers bottom-up instead of top-down, but he stresses the difficulty of creating useful keyword searches for doing this. (Nicholson 2013, 66–67) Adrian Bingham has also pointed this out, and has furthermore highlighted the danger that keyword searches (as well as other text mining techniques) pluck individual articles out of their original context, ignoring their position on the page, surrounding articles, and illustrations. (Bingham 2010, 230) Furthermore, Johanna Drucker has indicated that digital humanities scholars often aim to reduce complexity and remove ambiguity, while these are two values humanities research has to cherish, not avoid. (Drucker 2009, 5–7; Collini 2012, 65–84)

This paper takes such warnings into account and shows how these problems are being addressed by researchers working with the digitized newspaper database of the Dutch national library, thereby offering more concrete versions of the rather general solutions (e.g., ‘we should not forget the article’s context’) that are often suggested. At present, this database contains over 10 million pages from more than 200 newspapers and periodicals published between 1618 and 1995.³ It can be approached in two ways: through Texcavator (in development, not yet publicly available) and through the national library’s Delpher tool (publicly available).

The paper discusses a specific use case in which both tools are combined and used alongside traditional historical methods: researching identity formation in public discourse. It focuses on the identities of (health) risk groups, groups of people that are classified as ‘at risk’ with help of (health) risk factor classifications like the body mass index (BMI). For example, nowadays, people with a BMI above 25 are classified as ‘at risk’ because of their high body weight. This classification and the construction of this group is not a necessary outcome of biomedical research on the human body; instead it is historically contingent, strongly rooted in culture and practice. (Hacking 2007a, 2007b) The construction of these risk groups and the formation of their identity takes place for a significant part in public discourse. Digitized newspapers are valuable sources to study this identity formation: they provide a good entry into public discourse and typically span long time periods, enabling researchers to analyze the fluctuations in the identity of these groups (e.g., fluctuations between whether or not they are seen as (and see themselves as) ‘ill’).

The paper presents the first results of the investigation of the identity construction of the risk group ‘overweight people’ between 1890 and 1990. It focuses in particular on newspaper advertisements in the first part of this period — a choice based on distant reading of the corpus with help of Texcavator. The paper discusses how Texcavator and Delpher have been used, focusing in particular on the interaction between close and distant reading necessary to do this type of research. It shows how the direct connection between Texcavator and Delpher makes sure the researcher is constantly only one or two mouse clicks away from viewing the single articles in their original context — on the page, including illustrations, within the full issue of the periodical, as if going through newspapers on microfilm (or, depending on the size of the computer screen, leaving through them on broadsheet). Furthermore, it shows how Texcavator’s built-in visualization tools (time lines with number of articles diagrams, word clouds, named entity recognition) can be used to go back and forth between distant and close reading in order to build sophisticated queries that can easily be refined and modified within the tool.

In this way, the paper shows the challenges but also the new heuristic possibilities of doing historical research in digital repositories of newspapers.

Notes

1. www.translantis.nl

2. On Texcavator see Huijnen et al. 2013; on Delpher see www.delpher.nl. Delpher has replaced the earlier, no longer available tool Lucene.
3. delpher.kb.nl

References

- Bingham, A. (2010). *The Digitization of Newspaper Archives: Opportunities and Challenges for Historians*, Twentieth Century British History, 21: 225–231.
- Collini, S. (2012). *What Are Universities For?* London: Penguin.
- Drucker, J. (2009). *SpecLab: Digital Aesthetics and Projects in Speculative Computing*. Chicago: University of Chicago Press. Available at <<http://public.eblib.com/EBLPublic/PublicView.do?ptid=448540>>.
- Hacking, I. (2007a). Kinds of People: Moving Targets. *Proceedings of the British Academy*, 151: 285–318.
- Hacking, I. (2007b). Where Did the BMI Come From. In: *Bodies of Evidence: Fat Across Disciplines*. Cambridge: Centre for Research in the Arts, Social Sciences and Humanities, University of Cambridge.
- Huijnen, P., Laan, F., de Rijke, M., Pieters, T. (2013). A Digital Humanities Approach to the History of Science: Eugenics Revisited in Hidden Debates by means of Semantic Text Mining. In: Wierzbicki, A., Jatowt, A., Nadamoto, A., Leidner, J. (eds.), *SocInfo 2013 Workshop Proceedings, 1st International Workshop on Histoinformatics*, Cham et al: Springer.
- Nicholson, B. (2013). The Digital Turn. *Media History*, 19: 59–73.

Extracting Relationships from an Online Digital Archive about Post-War Queensland Architecture

Hunter, Jane

The University of Queensland, Australia

Macarthur, John

The University of Queensland, Australia

Van der Plaat, Deborah

The University of Queensland, Australia

Gosseye, Janina

The University of Queensland, Australia

Muys, Andrae

The University of Queensland, Australia

Macnamara, Craig

The University of Queensland, Australia

Bannerman, Gavin

The State Library of Queensland

The “Architectural Practice in Post-War Queensland:

Building and Interpreting an Oral History Archive” project is a collaboration between the University of Queensland, the State Library of Queensland (SLQ) and four of the longest-standing architectural firms in Queensland. The project’s aim is to build a comprehensive online multimedia digital archive that documents architectural practice in post-war Queensland (1945–1975) – a period that was highly significant in Queensland’s architectural history but that remains largely undocumented. The goal was to use innovative Semantic Web technologies to link tacit knowledge extracted from individual oral histories to tangible knowledge (drawings, books, photographs, manuscripts) that exists within personal archives, firm archives as well as State and institutional archives and libraries.

The approach involved firstly conducting and recording a series of oral history interviews and public forums with the key architects from this period. These events comprise both private

interviews, one-on-one conversations between the project team and architect/s as well as a number of larger public forums held at the SLQ that focus on a specific theme (education, style, climate, regionalism, etc.)

The oral history interviews and the public forums are filmed, captured as digital files (.wav and .avi) and transcribed. Both manual tagging and text processing tools are applied to the transcripts to semantically tag key entities (architects, firms, structures, places, dates) mentioned in the interviews and extract new knowledge in the form of RDF graphs. The resulting RDF graphs document relationships between architects, firms and buildings (with attribution to the source) and are able to be displayed, edited, saved and re-used via the LORE compound object authoring software (Figure 4).

This paper describes our approach to establishing the online archive and evolving knowledge-base¹ that together have been designed to be used for research, teaching and practice within the disciplines of history, architecture and design.

An overview of the system architecture is shown in Figure 1. The system uses the Omeka content management system to support the upload and description of content (oral history files, transcripts, photos, drawings, articles etc.) by the project collaborators. In addition the system provides the following components and functionality:

- An OWL Architecture ontology was defined that specifies the core classes, class hierarchy and properties associated with each class (Figure 1);
- D2RQ is used to convert Omeka metadata to RDF and save it to a Sesame RDF triple store with a SPARQL query interface;
- User-authenticated annotation tools enable users to semantically tag transcripts by identifying people, places, buildings, firms, and events mentioned in the interviews;
- The EYE N3 Semantic Reasoner (N3) is applied to the Sesame RDF triple store to reconcile common entities (via URLs) and to infer relationships between key entities (architects, firms, structures/buildings and articles/publications);
- A Search and Browse engine (based on Solr) enables users to search for specific entities or perform full-text searching across all transcripts and articles (and jump to the audio/video segments that contain the matching search term);
- Word clouds and word frequency histograms (generated from the oral history transcripts using D3) enable architectural historians to understand the main themes and influences on key architects from this period;
- Mapping and timeline interfaces enable users to interactively browse and retrieve information (interviews, photos, drawings) about buildings, people or events via maps and timelines;
- The LORE tool enables the visualization, editing, sharing and re-use of RDF graphs that document relationships between architects, firms, buildings, and related documents (Figure 3)

At the time of writing this abstract, the archive/database contains 64 interviews, of average length 83 mins. It also contains 64 transcripts, 725 photos, 612 articles, 305 line drawings and detailed information about 464 architects, 119 firms and 357 buildings/structures. The archive is growing continuously as more interviews and associated content are uploaded and annotated. The architectural historians involved in the project and their students, review the transcripts and using the integrated annotation tools identify and tag the names of people/architects, firms/organizations, buildings/structures and places. As new people, structures, firms and places are tagged/identified, they are added to the ontology. Authenticated users can also annotate relationships between people, between people and firms and between people and buildings, by drawing on a controlled vocabulary of relationship types. The reasoning engine then reasons across these relationships to infer new implicit relationships that can be recorded, searched and visualized through the LORE RDF graph visualization tool.

Architects who studied and worked in Queensland during the post-war period are also invited to register, login and submit their own details including a chronology of practice and to provide feedback to the existing content. An additional blog monitored by the project team encourages the broader

community (those outside the profession) to comment on aspects of post-war architecture (e.g., nominate their favourite building) and to upload related materials such as photographs or plans.

Future work plans include undertaking a detailed user evaluation of the system with a set of test users that comprises architectural historians from academic, government and industry as well as users from the local architectural community - and refining and extending the system based on user feedback.

Finally, our paper will also describe the challenges that this multi-disciplinary project faces including: how to attract and retain an active community of contributors; ensuring the archive's sustainability, resolving issues of identity resolution and implementing quality control over the community-generated content.

Biography:

Professor Jane Hunter is the Director of the eResearch Lab at the University of Qld – where she leads a team of post-docs, PhD students and software engineers working on innovative e-research services for a wide range of applications and communities. She has published over 100 peer-reviewed papers on semantic web, digital libraries and e-research and is currently the Deputy Chair of the Australasian Association for Digital Humanities and Chair of the Academy of Sciences Committee for Data in Science. She is a CI on the Mellon-funded Open Annotation Collaboration (OAC) project, the NeCTAR-funded HuNI and Aust-ESE projects and the ARC Linkage Project “Architectural Practice in Post-War Queensland: Building and Interpreting an Oral History Archive”.

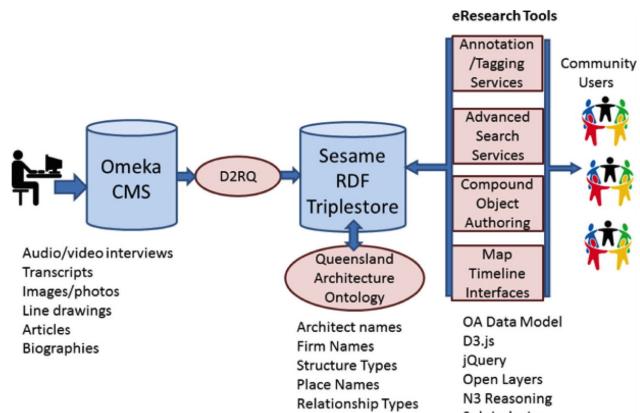


Fig. 1: Technical Components underlying the Post-War Queensland Architecture Knowledge Base

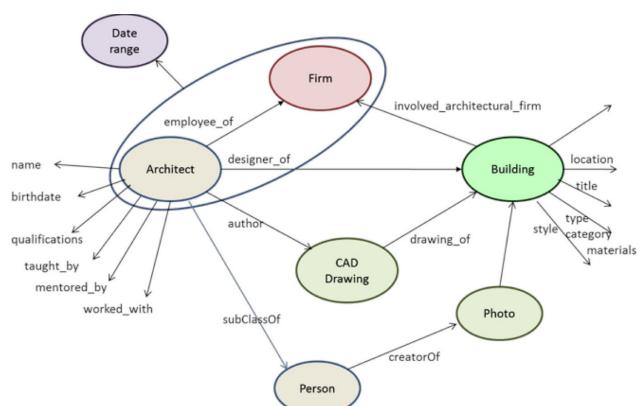


Fig. 2: Overview of the Ontology underlying the Post-War Qld Architecture Knowledge-base

DIGITAL STORIES



Fig. 3: Screen Shot of the Web Portal: Digital Archive of Queensland Architecture

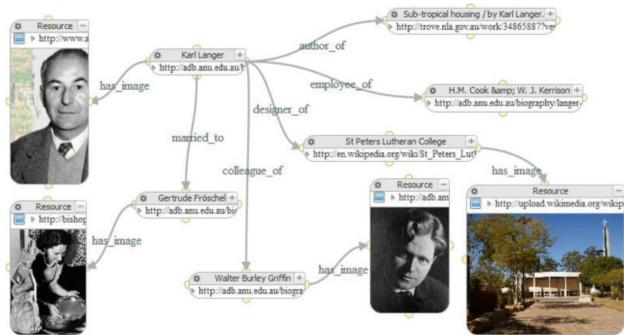


Fig. 4: LORE Visualization and Editing Interface to a Relationship Graph about Karl Langer

References

- Digital Archive of Queensland Architecture Web Portal* (2014). qldarch.net
- I (2014). Omeka. omeka.org
- Gerber A. and Hunter, J.** (2010). *Authoring, editing and visualizing compound objects for literary scholarship*. Journal of Digital Information, 11(1), 2010.
- Bizer, C. and Cyganiak, R.** (2014). *D2RQ Accessing Relationsl Databases as Virtual RDF Graphs*. d2rq.org
- Verborgh, R. (2011). *Semantic Reasoning with EYE*. n3.restdesc.org
- Bostock, M.** (2013). *D3 Data-Driven Documents* d3js.org

Student Collaborators in Digital Humanities Outreach and Advocacy: Strategies and Examples from the IDHMC at Texas A&M University

Ives, Maura
m-ives@tamu.edu
 Texas A&M University, United States of America

Earhart, Amy
 Texas A&M University, United States of America

Grumbach, Elizabeth
 Texas A&M University, United States of America

Mandell, Laura

Texas A&M University, United States of America

Our paper explores the role of student participants in outreach and advocacy for digital humanities centers. Although discussion of undergraduates in digital humanities generally focuses on classroom research activities (Croxall 2013), we define “student participant” broadly, including students who are employed by, or complete internships within, a digital humanities center. Because student participants are highly responsive to outreach efforts, and can themselves support outreach by serving as advocates for the center (and digital humanities scholarship in general) during their student years and beyond, cultivating student participants allows centers to enhance existing outreach efforts, so much so that the benefits for students and the center may justify refocusing energy and resources to better support student engagement. At the same time, digital humanities centers must address institutional barriers and ethical concerns that emerge when students are involved in research. Our talk will provide examples of the rationale and strategies that the IDHMC (Initiative for Digital Humanities, Media and Culture) is developing to strengthen student involvement in digital scholarship. Recognizing that funding and other factors create obstacles to student involvement, we are working towards a flexible model that can be adapted to suit local circumstances.

Although discussions of “outreach” for the digital humanities often focus on publicizing projects (see Brennan 2013), centers are expected to perform generalized outreach to promote the center and its work and to make research, tools, and training available to academic and general audiences (Nowviskie 2011). While these forms of outreach are important, they create considerable demands upon staff (Nowviskie 2010; Ramsay 2010) and are limited in their reach and effectiveness. Bringing traditional humanities scholars into digital humanities scholarship is time and labor intensive, and projects and tools often target specific audiences, leaving out constituencies that are institutionally or strategically important for the center’s growth. Student participants both supplement and extend these forms of outreach.

Focusing outreach efforts on student participants may be more effective than targeting humanities faculty who are not already engaged in digital scholarship. Unlike faculty, whose interest in new technologies is constrained by multiple factors (Wiberly and Jones 2000; O’Donnell 2009), students have time to invest in learning new technologies and approaches and are less invested in traditional disciplinary paradigms. Before graduation, student participants enhance outreach by bridging disciplinary and institutional divisions and communicating their digital humanities interest to other students and faculty. After graduation, student participants may still contribute by incorporating digital tools and collaborative practices into their professional activities, serving as intermediaries between the center and external partners, and continuing to participate in scholarship through crowdsourcing or other means.

The IDHMC uses grant funding to support student involvement in the classroom and as employees, and is hoping to secure funding for a pilot internship program. In 2012, through an internal (Tier One Program) grant, Laura Mandell developed an interdisciplinary, stacked (undergraduate and graduate) course that involved students in designing the IDHMC’s Humanities Visualization Space (HVS). That course has now become a permanent offering. But to reach more students from multiple disciplines, the IDHMC invests in student learning beyond targeted digital humanities classes. Mandell’s eMOP (Early Modern OCR Project) Mellon grant includes graduate and undergraduate student workers who have made important contributions to the project and to the digital humanities community (Torabi, Durgan and Tarpley 2013). Most recently, Mandell, Ives and Earhart targeted multiple forms of student engagement in an internal grant in support of the development of ARC (Advanced Research Consortium, idhmc.tamu.edu/projects/arc/) and the HVS. To infuse digital humanities scholarship in the wider curriculum, this grant funds the creation of workshops and class activities that faculty can integrate into existing classes, as well as a research competition for students who have used the ARC data and the HVS in their projects (we will provide links to these

materials in our presentation). Finally, IDHMC is piloting an unpaid undergraduate internship program (idhmc.tamu.edu/blog/2013/06/21/idhmc-undergraduate-internships/) through which students can work on any IDHMC affiliated project. Through employment and internships, we are drawing students from departments that do not offer DH coursework or research opportunities, increasing our impact on campus while creating valuable learning experiences for a variety of students.

To pursue these goals, we continually navigate institutional and disciplinary barriers, including faculty resistance, administrative and curricular roadblocks, and funding issues. The difficulties inherent in adding new courses to the official curriculum make other means of outreach to students especially attractive and valuable. We hope that we will be able to offer stipends to future interns, perhaps ultimately creating something like DePauw's ITAP (Information Technology Associates Program), an internship program in which liberal arts undergraduates are given a year of technical training and then placed in academic or administrative internships. Programs such as ITAP demonstrate that incorporating undergraduate students into digital humanities research and outreach is a strategy that is neither limited to digital humanities centers such as IDHMC or to research universities, nor dependent upon external grant funding (though there is no doubt that the necessity of finding some source of funding for student researchers is a challenge across the board). Such a strategy does, however, require acceptance of a capacious definition of what "counts" as digital humanities, and resistance to the idea that a valuable student experience can only be obtained through exposure to the kinds of digital humanities work that require heavy institutional investment in equipment and staff support. Ideally, students working at IDHMC will experience a variety of modes of digital humanities research, from digital archives (such as the Cervantes Project), to tool development (via the OCR testing and development that is part of eMOP), to academic instances of social media (the Anachronaut Facebook game for OCR correction), to humanities data visualization. Students working at smaller institutions with fewer resources will not have as many options as we do, but the necessity of limiting hands-on experience to certain kinds of projects does not prevent faculty from exploring other forms of digital scholarship through other means, including collaborative arrangements with other institutions that offer different digital humanities emphases. Regional or national partnerships that facilitate digital project information sharing (something as simple as having faculty involved in a project Skype in to a class) are one way to expand student involvement beyond a single institution's resources and expertise; over time, more intensive collaborations – a "spring break" research experience at another institution, for example -- might be developed. More important still is providing the kind of digital humanities acculturation that Geoffrey Rockwell and Stefan Sinclair describe, which students learn to "work in interdisciplinary teams, apply digital practices to the humanities, manage projects, [and] explain technology and build community" (Rockwell and Sinclair 2013). The goal of acculturation is one that can be pursued in many ways and across many institutional structures. For the purposes of outreach, it is extremely valuable given its relevance to students' career paths.

While we work to develop a funded internship program, we offer independent studies to all interns so that they receive formal academic credit for their work. As a general practice, we encourage all affiliated students to present or publish their work and we include them in our own presentations and publications (Ives et al; Earhart; Mandell 2013). We also feature student contributors in publicity for IDHMC and its projects. While many of these practices are (or should be) universal, articulating them is important, both to insure a positive experience for students and to equip them to serve as advocates for IDHMC and digital humanities generally.

The overall goal of our work with students is to enact the values of digital scholarship formulated by Lisa Spiro: openness, collaboration, collegiality and connectedness, diversity, and experimentation (Spiro 2012). We want to instill an ethical perspective that is both specific to digital humanities and generalizable to students' future careers. Expanding upon studies that emphasize collaboration and cross-disciplinary

learning within the classroom (Norcia 2008), we aim at, in Julia Flanders' words, fostering "a professional academic ecology" that illuminates "the diversity of working roles that contribute to the production of scholarship" (Flanders 307). Our focus on multiple forms of student participation allows us to define the ethos of collaboration broadly, emphasizing respect for and recognition of contributions across disciplinary boundaries as well as educational/professional roles (student, employee, faculty, staff). Successful outreach can be defined in many ways, but the most meaningful measure may well be the extent to which students carry this ethos with them when they leave IDHMC.

References

- Brennan, Sheila.** (2013). "Four P's of Digital Project Outreach." Lot 49. 1 August 2013. www.lotfortynine.org/2013/08/four-ps-of-digital-project-outreach/
- Croxall, Brian, et al.** (2013) "The Future of Undergraduate Digital Humanities." 17 July 2013. dh2013.unl.edu/abstracts/ab-292.html
- Earhart, Amy.** "Alex Haley's Malcolm X: 'The Malcolm X I Knew' and notecards from The Autobiography of Malcolm X," an introduction and edition, forthcoming in Scholarly Editing.
- Flanders, Julia.** (2012) "Time, Labor, and "Alternate Careers in Digital Humanities Knowledge Work." Debates in the Digital Humanities. Ed. Matthew Gold. Minneapolis: University of Minnesota P. 292-308
- IDHMC Undergraduate Internships* (21 June 2013). idhmc.tamu.edu/blog/2013/06/21/idhmc-undergraduate-internships
- Information Technology Associated Program.* www.depauw.edu/it/ita
- Ives, Maura; Del Hierro, Victor; Kelsey, Bailey; Smith, Laura Catherine and Christina Sumners.** (2013) "Encoding the Discipline: English Graduate Student Reflections on Working with TEI." Journal of the Text Encoding Initiative: Selected Papers from the 2012 TEI Conference. Issue 6, December 2013. jtei.revues.org/88
- Mandell, Laura.** (2013). *Rev. of Debates in Digital Humanities*, ed. Matthew K. Gold, in Information and Culture: A Journal of History, with Matthew Davis, Tess Habersetz, Jacob Heil, Shawn Moore, Laura Perrings, and Katayoun Torabi. 19 March 2013. www.infoculturejournal.org/book_reviews/idhmc_gold_DebatesDH
- Mandell, Laura, et al.** (2012) *Mellon Foundation Grant Proposal: "OCR'ing Early Modern Texts."* 30 June 2012. <<http://idhmc.tamu.edu/projects/Mellon/eMOPPublic.pdf>>.
- Mandell, Laura, Maura Ives, and Amy Earhart.** (2013) "ARC: Research and Student Engagement in the Digital Humanities." Strategic Development Fund Proposal, TAMU. 1 April 2013.
- Nowviskie, B.** (2011) "Are You our New Head of Outreach and Consulting?" Scholar's Lab blog. 23 February 2011. www.scholarslab.org/announcements/head-of-outreach-consulting/
- Nowviskie, B.** (2010) "Eternal September of the digital humanities." 15 October 2010. nowviskie.org/2010/eternal-september-of-the-digital-humanities
- eMOP (Early Modern OCR Project). emop.tamu.edu .
- Norcia, M.** "Out of the Ivory Tower Endlessly Rocking: Collaborating across Disciplines and Professions to Promote Student Learning in the Digital Archive." *Pedagogy* 8(1): 91-114. muse.jhu.edu/journals/pedagogy/v008/8.1norcia.pdf .
- O'Donnell, James.** (2009) "Engaging the Humanities: The Digital Humanities." *Daedalus* 138.1 Winter 2009: 99-104.
- Ramsay, Stephen.** (2010) "Centers of Attention." 27 April 2010. dh2013.unl.edu/abstracts/ab-292.html
- Rockwell, Geoffrey and Stefan Sinclair** (2012). "Acculturation and the Digital Humanities Community." *Digital Humanities Pedagogy: Practices, Principles and Politics*. Ed. Brett D. Hirsch. Open Book Publishers. openbookpublishers.com/htmlreader/DHP/chap07.html
- Spiro, Lisa** (2012). "This is Why we Fight: Defining the Values of the Digital Humanities." Debates in the Digital

Humanities. Ed. Matthew Gold. Minneapolis: University of Minnesota P. 16-35. dhdebates.gc.cuny.edu/
Stephen E. Wiberley, Jr. and William G. Jones (2000). "Time and Technology: A Decade-Long Look at Humanists' Use of Electronic Information Technology," College & Research Libraries 61 (Sept. 2000): 421-431.

Torabi, Katayoun, Jessica Durgan, and Bryan Tarpley (2013). "Early Modern OCR Project (eMOP) at Texas A&M University: Using Aletheia to Train Tesseract." Paper presented at DocENG 2013, 11 September 2013. www.doceng2013.org/programme

Using Social Network Analysis to Reveal Unseen Relationships in Medieval Scotland

Jackson, Cornell Alexander
 Kings College London, United Kingdom

Introduction

This short paper will describe the on-going research being conducted jointly by Kings College London and the University of Glasgow to understand the social networks of the medieval Scottish elites from the years from 1093 to 1286. This paper will start with a description of social networks and the concepts of social network analysis. It will then move on to describe some of the uses social network analysis has been put to in historical research. This will be followed by a description of the People of Medieval Scotland database which provides the data for this research. Finally, the social network analysis techniques used in this research will be described and the preliminary results will be discussed.

Social Network Analysis

Social networks are defined and measured as connections among people, organisations, political entities (states and nations) and/or other units. Social network analysis is a theoretical perspective and a set of techniques used to understand these relationships (Valente 2010, pg. 3). Christakis and Fowler (2010, pg. 32) say that the science of social networks provides a distinct ways of seeing the world because it is about individuals and groups and how the former becomes latter.

Valente (2010, pgs. 3 – 7) says that relationships matter because relationships influence a person's behaviour above and beyond the influence of his or her attributes. A person's attributes does influence who people know and spend time with: their social network. Valente quotes Borgatti et al (2009), "one of the most potent ideas in the social sciences is the notion that individuals are embedded in thick webs of social relations and interactions". The reason that social networks are so important is because human beings are ultra-social animals that create social networks (Haidt, 2006). Christakis and Fowler (2010, pg. 214) add that human beings just don't live in groups, they live in networks. Valente argues the traditional social science approach of using random sampling is not adequate for measuring network concepts because random sampling removes individuals from the social context that may influence their behaviour. Valente explains that one primary reason social network research has grown in recent decades is that scholars have become dissatisfied with attributes theories of behaviour. Many attribute theories have not explained why some people do things (e.g. quit smoking) while others do not. Social network explanations have provided good explanations in these cases.

Use of Social Network Analysis in History

The seminal work in using social network analysis in historical research is Padgett and Ansell's (1993) research on the rise of the Medici in renaissance Florence. Their work showed that the rise of the Medicis came from their ability, especially the ability of Cosimo de Medici, to take advantage of the gaps in connections in the social network which the Medicis were able to bridge to take political control of Florence. Since then, the use of social network analysis in historical research has been steadily increasing.

Using Social Network Analysis with the People of Medieval Scotland Database

The People of Medieval Scotland (PoMS) database holds data on all known people between 1093 and 1314 mentioned in over 8600 contemporary documents. This was funded by the Arts and Humanities Research Council in the United Kingdom. The current research is part of the Transformation of Gaelic Scotland project funded by the Leverhulme Trust. This exploratory research has the goal of understanding the role of social networks among the elite of medieval Scotland. It also has the goal of exploring the appropriateness of social network analysis techniques for this data set, and perhaps for other similar collections

The first technique used was 2 mode networks. In 2 mode networks, two sets of actors are dealt with. This method comes from the pioneering work of Davis et al (1941). In this research, the two sets of actors are legal documents called charters and the people who witness them. As a result, you will see links between witnesses and charters but not among the witnesses and charters. This becomes even more useful by the affiliation technique. Here the software is asked to create a 1 mode network or a network with only one set of actors by connecting witnesses who have witnessed the same charter together. The software can also keep track of how many times a particular witness has witnessed a charter with every other witness. The theory is that the more often two witnesses witness charters together, the more probable there is an actual social relationship between the two people. Therefore, as the number of charters witnessed together rises, the more probable the resulting network is an actual map of the social relationships.

Other techniques used include:

- 2 mode network with witnesses by locations to identify geographically clustered witnesses
- Ego networks where the focus is on all the people connected to a selected person and the interconnections among these people
- Directed network of grantors and beneficiaries. This network is directed because the direction is always from the grantor to the beneficiary.
- Using cluster analysis and structural equivalence to see if witnesses can be clustered by the similarity of their network connections
- Using network models of diffusion of innovations to track how charter innovations spread

Findings so far

This work is still very preliminary but some interesting findings have appeared from the use of social network analysis. One good example of this is Duncan II, Earl of Fife. Historians knew he was a very prominent noble in Scotland but social network analysis revealed a possible further role he played in Scotland.

Duncan has witnessed more than 20 charters with 27 people while William del Bois, the chancellor whose role is to manage charters, has only done that with 15 other witnesses. Also, Duncan has witnessed more than 40 charters with 7 people while William has done the same with only 2 other witnesses. However, Duncan has witnessed charters with 630 other witnesses while William, the chancellor, has witnessed charters with 479 other witnesses. Overall, William has witnessed 213

charters while Duncan has witnessed 210 charters. So the question is why does Duncan, Earl of Fife have so many more connections than anyone else? There is no definitive answer to this question yet but the leading hypotheses centre on Duncan's possible role in the government, taking advantage of his brokerage opportunities and enhanced social skills.

The grantor/beneficiary network showed that those who gave the most grants were kings, popes, bishops and senior noblemen such as earls. Those who received the most grants were mainly ecclesiastical institutions such as abbeys, priories and cathedrals. In addition personas such as Saint Cuthbert and the Blessed Virgin Mary received large number of grants. However, none of this was very surprising given the nature of medieval Scottish society.

Ego networks for a number of people have been generated. We have compared these ego networks by density, brokerage opportunities and how often the ego acts as a bridge inside the network. No general trends have been discovered yet. We have also looked at turning Burt's (1992) work on its head by looking not at brokerage opportunities but at the interconnections to determine the characteristics of their social networks. As of now, this has not been completely successful.

It is too soon at this date to report findings on using network models on the diffusions of innovations. However, we see this as an exciting prospect as it will allow the tracking of how charter innovations spread or did not spread throughout medieval Scotland. Historians have identified several charter innovations to investigate and we hope to report on this at the conference.

The 2 mode witness by location network did not work because of bad location data in the database which is now being corrected. But, the biggest data issue in using social network analysis with this data is that the People of Medieval Scotland database has only legal documents in it and does not have the marriage, baptismal and tax records that Padgett and Ansell (1993) used. These additional records would allow us to confirm relationships that can be inferred especially from the 2 mode network analysis.

Summary

In summary, while this research is still preliminary, it has shown the power of social network analysis to bring a new perspective to old data. Duncan II, Earl of Fife is an example of this. While the historians knew Duncan was very prominent in his time, they had no idea that he might have a possible role in running the Scottish government during the reign of William I. The use of network models of the diffusion of innovations to see how charter innovations did or did not spread in medieval Scotland is another example where this technique will allow us to show the mechanism of how these innovations spread.

References

- Borgatti, S. P. and A. Mehra, D.J. Brass and G. Labianca** (2009) "Network Analysis in the Social Sciences", Science, Volume 323, pp. 892 – 895
- Burt, R.** (1992) *Structural Holes: The Social Structure of Competition* London: Harvard University Press.
- Christakis, Nicholas and James Fowler** (2010) *Connected: The Amazing Power of Social Networks and How They Shape our Lives*. London: Harper Press
- Davis, Allison, Burleigh B. Gardner and Mary R. Gardner** (1941) *Deep South: A Social Anthropological Study of Caste and Class* Chicago: University of Chicago Press
- Haidt, Jonathan** (2006) *The Happiness Hypothesis: Finding Modern Truth in Ancient Wisdom*. New York: Basic Books
- Padgett, John and Christopher Ansell** (1993) "Robust Action and the Rise of the Medici, 1400 – 1434", *American Journal of Sociology* Volume 98, Issue 6, pp. 1259 – 1319
- Valente, Thomas W.** (2010) *Social Networks and Health: Models, Methods and Applications*. Oxford: Oxford University Press

IMPACT : un dispositif de transcription et de commentaire de l'oral, pour l'enseignement et la recherche

Jacquin, Jérôme

jerome.jacquin@unil.ch

Université de Lausanne

Gradoux, Xavier

xavier.gradoux@unil.ch

Université de Lausanne

Cette contribution vise à présenter IMPACT (Interface Multimédia: Présentation, Analyse, CommenTaire), un **dispositif technopédagogique hybride**, c'est-à-dire mêlant apprentissage en présence et à distance⁽¹⁾ et visant l'amélioration du scénario pédagogique initial par l'usage de Technologies de l'Information et de la Communication⁽²⁾. Plus généralement, la contribution entend situer le développement et l'utilisation de cette technologie dans une réflexion plus générale, celle des Humanités Digitales, sous l'angle spécifique du **rapport entre enseignement et recherche dans la constitution progressive de bases de données enrichies**.

1. Besoins initiaux

L'enseignement qui a bénéficié du développement du dispositif s'inscrit dans le domaine des **Sciences du langage** et entend introduire les étudiants à l'analyse de conversations quotidiennes^{(3) ; (4) ; (5)}. En tant que champ disciplinaire, l'analyse conversationnelle insiste sur le **caractère mutuellement constitutif de l'analyse des données** (les enregistrements tirés de conversations authentiques) et de la **production du discours théorique** (l'abstraction et la systématisation de régularités). L'analyse conversationnelle se veut éminemment **descriptive et inductive**, privilégiant un **contact direct et constant aux données** ainsi qu'une pratique de l'analyse attentive aux détails. Données et théorie apparaissent interdépendantes et il ne s'agit pas de négliger les premières au profit de la seconde. L'enseignement en question est semestriel et articule une partie "cours", où les enseignants sensibilisent les étudiants aux outils d'analyse en mobilisant de nombreux exemples, et une partie "travaux pratiques" à l'occasion de laquelle les étudiants s'approprient la matière en effectuant un enregistrement, en le transcrivant finement et en proposant une analyse.

L'obstacle majeur auquel était confronté le scénario pédagogique concerné résidait dans la **manipulation constante et fastidieuse de sources hétérogènes et dispersées** (sources audio-visuelles, transcription d'extraits, analyses de ces extraits, documents théoriques), mais nécessaires et complémentaires. L'utilisation du CMS Moodle a certes permis de rassembler les fichiers sur un serveur centralisé et a favorisé l'émergence d'un apprentissage à distance. Elle n'a néanmoins pas facilité l'intégration des données : les différentes sources, dont la consultation nécessite à chaque fois un téléchargement, ne peuvent être traitées de manière synchronisée en raison des différences de formats et de types de fichiers. Ceci ralentit le processus, nuit à l'attention et donc rend le rapport aux données opaque sans gain réflexif ou pédagogique⁽⁶⁾.

2. Solutions trouvées

Après avoir vérifié leurs intuitions sur un prototype d'interface rapidement assemblé par les ingénieurs pédagogiques de leur institution, les enseignants ont déposé un projet au fond d'innovation pédagogique de l'Université de Lausanne (www.unil.ch/fip). Il a abouti, en 2011, à la confection d'une **interface multimédia permettant de construire des fiches qui rassemblent dans un même espace de consultation**

et d'édition différentes ressources hétérogènes, telles qu'une source audiovisuelle, sa transcription et son analyse. Éditable (par les enseignants et les étudiants) et visionnable (par internet ou par vidéoprojection) de partout et à tout moment⁽⁷⁾, l'interface a apporté une plus-value ergonomique indéniable et appréciée par les utilisateurs.

The screenshot shows a window titled "21. Marques et stratégies d'introduction". It includes a video player at the top showing two men in a debate, with the time 52:24 and 53:09. Below the video is a transcription of their conversation. The transcription starts with "Contexte : après 2h du débat de l'entre-deux-tours entre Nicolas Sarkozy (NS) et Ségolène Royal (SR), diffusé sur TF1 et France 2 en direct le 2 mai 2007. Doublette : Arlette Chabot (AC) et Patrick Poivre d'Arvor (PPD)". The transcription text is as follows:

1 NS et la certitude de la réalisabilité de cette promesse (...) ça s'ira
2 le droit opposable (...) et la capacité [mala] (...) est-ce-
3 SR nd'aller devant un tribunal pour dire que je suis [droits]
4 NS «vous voyez n'admettez (...) c'est ni ridicule (...) ni malveillant [qui] peut-
5 SR (...) c'est peut-être même ce qui fait la différence (...) entre
6 la vieille politique (...) et la politique [général]
7 PPD réussis tous les deux [une promesse] (...) j'ai quelque chose à
8 dire (...) parce que là je pense (...) je pense
9 AC [un mot puis euh et on enchaîne]
10 SR je suis scandalisée de ce que je viens d'entendre (...) parce que
11 AC je joue (...) avec le handicap (...) comme vous venez de le faire
12 SR je suis très professionnelle mais pourtant lorsque j'arrive à l'heure
13 AC j'étais ministre de l'enseignement scolaire (...) c'est MOI qui ai
14 AC [un mot puis euh et on enchaîne]
15 SR je suis scandalisée de ce que je viens d'entendre (...) parce que
16 AC je joue (...) avec le handicap (...) comme vous venez de le faire
17 SR je suis très professionnelle mais pourtant lorsque j'arrive à l'heure
18 AC j'étais ministre de l'enseignement scolaire (...) c'est MOI qui ai
19 AC [un mot puis euh et on enchaîne]
20 SR je suis scandalisée de ce que je viens d'entendre (...) parce que
21 AC je joue (...) avec le handicap (...) comme vous venez de le faire

Below the transcription is a section titled "Théorie" with a note about the emergence of new topics through the combination of linguistic markers. A note also states that the topic has been enriched by interaction. The bottom of the window shows a note about the creation of pre-requisites for marking, followed by a note about exemplification.

Fig. 1: Exemple de fiche générée par IMPACT.

Plus concrètement, IMPACT permet aux étudiants et aux enseignants de :

- Diffuser une source audio ou audio-vidéo dans son intégralité ou en en choisissant un extrait.
- Piloter la lecture de la source au clavier (lecture/pause; ralenti; retour de 2 sec.; rembobinage rapide).
- Transcrire la source ou un extrait en respectant les standards du domaine mais dans un canevas intuitif et ergonomique.
- Exporter la transcription ou en importer une existante.
- Synchroniser ou non le défilement de la transcription et de la source.
- Ajouter des commentaires textuels (théoriques ou analytiques).
- Joindre des documents PDF ou des images.
- Introduire un scénario pédagogique grâce à une logique d'étapes, permettant à l'enseignant d'évaluer progressivement le travail de l'étudiant.
- Coordonner chaque étape à un rendu de devoirs dans Moodle ou un message aux enseignants.

Pour davantage de détails, voir⁸.

3. Evaluation pédagogique et impact pour la recherche

De manière réflexive, l'élaboration d'une telle interface a également (i) enrichi le scénario pédagogique par la réalisation de fiches d'enseignement (la matière du cours est segmentée en unités thématiques cohérentes, articulant théorie et pratique) et de fiches d'étudiants (mise à disposition de la plate-forme pour présenter le travail de recherche personnel), (ii) favorisé une relative autonomisation, saluée par les étudiants, de l'enseignement vis-à-vis de la situation de co-présence (cette autonomisation serait compatible, à terme, avec un apprentissage à distance) et (iii) influencé positivement la motivation et l'engagement des étudiants.

Dans cette perspective, une enquête a été menée auprès des étudiants afin de déterminer dans quelle mesure l'utilisation du dispositif pour s'approprier les notions et réaliser les travaux avait modifié leur apprentissage et leur façon d'appréhender la matière. L'étude révèle, entre autres choses, que les attentes des étudiants envers l'enseignement en présence sont modifiées par l'utilisation d'IMPACT, dans le sens où l'exemplification et la problématisation des notions prennent le pas sur leur seule présentation.

De l'utilisation croissante d'IMPACT dans le domaine de spécialisation – les Sciences du langage – a en outre émergé une réflexion plus générale sur l'élaboration et l'interrogation de bases de données orales comme enjeu pédagogique.

Quels sont les défis pédagogiques auxquels l'oralité est actuellement confrontée ? Les **Humanités Digitales** sont-elles prêtes à intégrer l'oralité, tant sous l'angle de la recherche que de l'enseignement ? Comment se donner les moyens de rendre compte de l'**hybridation progressive des dispositifs de communication contemporains**, qui articulent de plus en plus finement oralité et scripturalité⁽⁹⁾ ?

Partant de ces trois questions, la comparaison des travaux réalisés dans IMPACT avec les **principales bases de données orales francophones** (VALIBEL, ESLO, CLAPI) s'avère intéressante, lorsqu'on considère ces différents outils sous l'angle spécifique de leur applicabilité pédagogique et particulièrement de la place qu'elles donnent à l'étudiantE. En outre, l'expérience d'IMPACT invite à voir dans quelle mesure un dispositif technopédagogique peut être adapté pour des usages émergents et inattendus, dans le cas présent la constitution progressive et collaborative, par les enseignants et les étudiants, de bases de données orales enrichies et disponibles pour la recherche.

References

- Charlier, B., Deschryver, N., & Peraya, D. (2006).** *Apprendre en présence et à distance : Une définition des dispositifs hybrides*. Distances et savoirs, 4, 469–496.
- Karsenti, T., & Larose, F. (Eds.). (2001).** *Les TIC... au cœur des pédagogies universitaires*. Québec: Presses de l'Université du Québec.
- Sacks, H., Schegloff, E. A., & Jefferson, G. (1974).** *A Simplest Systematics for the Organization of Turn-Taking for Conversation*. Language, 50(4), 696–735.
- Ten Have, P. (2007).** *Doing Conversation Analysis*. London: Sage.
- Hutchby, I., & Wooffitt, R. (2008).** *Conversation Analysis*. Cambridge: Polity Press.
- Karsenti, T., & Larose, F. (Eds.). (2001).** *Les TIC... au cœur des pédagogies universitaires*. Québec: Presses de l'Université du Québec.
- Alessi, S. M., & Trollip, S. R. (2001).** *Multimedia for learning: methods and development*. Boston: Allyn & Bacon.
- Gradoux, Xavier & Jacquin, Jérôme**, 2013, *Le projet IMPACT*, disponible à l'adresse www.unil.ch/impact, accédé le 21 octobre 2013.
- Lievrouw, L. A., & Livingstone, S. (2006).** *Handbook of new media: social shaping and social consequences - fully revised student edition*. London: Sage.

Digital learning in an undergraduate context: promoting long term student-faculty (and community) collaboration in the Susquehanna Valley, PA

Jakacki, Diane

diane.jakacki@bucknell.edu
Bucknell University

Faul, Katherine

faulk@bucknell.edu
Bucknell University

In this paper, we will present a case study of how an ongoing, multi-faculty, interdisciplinary DH project focused on the Susquehanna Valley in Pennsylvania has created, and continues to explore, ways in which students can excel both inside the classroom and outside. These DH projects involve undergraduates working with faculty on an unfolding expansive research project that affords otherwise unachievable opportunities for undergraduate student engagement, the development of new skills, and meaningful ongoing interaction

between the institution and community that have, in turn, furthered the scope and scale of the project.

While it is recognized that the most compelling pedagogical experiences bridge the divide between semesters, and even years, of study at the undergraduate level, there has been little examination to date of how digital pedagogy affords particularly effective forms of promoting long term student engagement, challenging students and instructor to consider and reconsider course matter from new and provocative vantage points. In early digital humanities programs, considerations of ensuring that “the acculturation and professionalization that takes place in the learning community is relevant to the students” has been largely situated in graduate programs, leaving undergraduates in learning environments where digital engagement focuses on tool training rather than one in which they learn digital “habits of mind” that involve participation in nuanced humanistic discourse with their professors.

In small liberal arts colleges, where close faculty/student interaction is at the core of high impact practices, opportunities to advance prolonged faculty-student collaboration can produce exceptional results. The Andrew W. Mellon Foundation has identified the importance of these opportunities for curricular digital engagement at liberal arts education through strategic multi-year digital initiative grants at an increasing number of liberal arts institutions.

A particularly valuable area for ongoing pedagogical engagement is in developing place-based projects that also enlist local communities in the digital, conceptual mapping of historical and cultural resources. And yet, such rich and nuanced considerations of local place, culture, and environment call for extended student engagement over time and even across years of undergraduate study. Thus, the traditional classroom model for the execution of such DH place based projects is inadequate. Extending the classroom outside (both spatially and temporally) allows for the development of rich deep local knowledge in both digital learning tools and content. Indeed, extending the faculty-student collaboration to include students from outside traditional humanities departments also reifies the value of interdisciplinary research at an early level and reflects the professional digital humanities research model employed by larger scale projects. Also, undergraduates engaged in digital humanities work can perform a sort of outreach, demonstrating that “the humanities [belong] to everyone, not just trained professionals” and thereby help “bridge the widening gap between academic humanities and broader American culture.”

At Bucknell University, beginning in 2011, faculty in Comparative Humanities, English, Geography, and Environmental Studies have developed and taught a slate of courses relating to issues vital to the interpretation and conservation of the environmentally impaired Susquehanna River. These courses form a de facto core curriculum designed around the region and consider questions of the environmental effects on regional resources, the eradication of the traces of Native American history and culture as a result of European immigration and settlement, and economic under-investment in post-industrial rural towns. To this point, DH engagement has focused on the collection and analysis of GIS-related materials, work that has been instrumental in the garnering of Federal recognition of the cultural importance of the Susquehanna River through its designation as a National Historic Trail under the umbrella of the National Parks Service. Students and faculty continue to work with non-profit agencies and the NPS in the development of digital layers of scientific and economic data that will expand the reach of this originally DH project.

A new phase of this project that will begin this summer with the digitization, transcription, and critical analysis of the collected correspondence, journals, records, and incidental papers of James Merrill Linn (1833-1897), held in the Archives of Bucknell University. This phase will involve Bucknell faculty, staff, and students in a research project that will develop within and beyond the classroom. The Linn Papers project is ambitious and represents a new model for collaborative digital humanities research and teaching at Bucknell. Because of Linn’s relationship with “place” (Bucknell, Lewisburg, and the sites of battles and campaigns in the Middle Atlantic and Southern states during the Civil War) this project offers an

important opportunity to expand and reconsider Bucknell’s commitment to considerations of “spatial thinking.” The Linn Papers archive includes documents in several media forms: manuscripts, drawings and sketches, printed records, archival newspaper clippings, and hand-drawn maps. Because of the multiple forms of inter-reliant media, the collection encourages analysis of people and places across document types. This form of analysis is best-suited to digital forms of curation and publication.

The Linn Papers project will a) make available in digital form a wealth of information about a historically under-resourced area of the Susquehanna Valley; b) teach students the principles and techniques of DH, and in particular, TEI-compliant XML; c) enable students to be active and engaged participants in the reframing of humanistic pedagogy and relevance in an age that sees almost daily public media questioning of the value of the humanities. This phase of the project will begin with a pilot undertaken in summer 2014 with the digitization of a selection of the papers. The Linn project will become a central facet of newly designed HUMN 100 courses offered in 2014-15 through the Comparative Humanities program, that are open to first and second year students only. Student engagement with the materials will also serve as a test case to determine best practices for incorporating TEI, GIS and network analysis skill development in a variety of courses, effectively creating a DH training stream at the university.

References

Rockwell, Geoffrey and Stéfan Sinclair (2012). *Acculturation and the Digital Humanities Community: Digital Humanities Pedagogy: Practices, Principles and Politics*. Brett D. Hirsch, ed. Cambridge: UK, Open Book Publishers. 177

Clement, Tanya. *Multiliteracies in the Undergraduate Digital Humanities Curriculum: Skills, Principles, and Habits of Mind*. *Digital Humanities Pedagogy: Practices, Principles and Politics*. Brett D. Hirsch, ed. Cambridge: UK, Open Book Publishers, 2012. 371, 384

In its 2012 annual report, Mellon identifies digital technology and pedagogy as key components in its commitment to assist liberal arts colleges find “a way for a limited number of faculty to teach the breadth and depth of a 21st-century curriculum and simultaneously help students develop their critical, creative, and intellectual capacities.” Bucknell is one of eighteen liberal arts colleges to receive one of these multi-year digital scholarship grants, and is working with its grant-funded peers to identify and implement learning-centric courses and projects. The Andrew W. Mellon Foundation Annual Report. 2012. 15. www.mellon.org/news_publications/annual-reports-essays/annual-reports/content2012.pdf

To develop the What Jane Saw project at the University of Texas at Austin Janine Barchas involved a number of undergraduate student technology assistants, summer research apprentices from Architecture as well as English, as well as undergraduates in her Austin course. Barchas, Janine. “Digitally Reconstructing the Reynolds Retrospective Attended by Jane Austen in 1813: A Report on e-Work-in-progress.” ABO Interactive Journal. March 1, 2012. www.aphrabehn.org/ABO/?p=1245#more-1245

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. Matthew K. Gold, ed. University of Minnesota Press.

Bucknell Helps Susquehanna River Secure Historic Designation. Bucknell University website. June 11, 2012. www.bucknell.edu/x77947.xml

The Rowling Case: A Proposed Standard Analytic Protocol for Authorship Questions

Juola, Patrick

juola@mathcs.duq.edu

Duquesne University, Juola & Associates

1. Introduction

One of the marks of a "mature science"¹ is the development of "standards" of analytic practice, based on shared "key theories, instruments, values, and metaphysical assumptions"² that scholars work with. This concept has been incorporated into US Law as a mark of reliable evidence.³ One of the weaknesses of authorship attribution is the absence of such standards of practice. For example, fifteen years ago Rudman estimated⁴ that more than 1000 different feature sets had been proposed for this task. This of course creates controversy about the appropriateness of methods and even the possibility of cherry-picking feature sets to a specific task to get a desired answer.

The solution demanded by Daubert is the use of a specific analytic technique, with standards controlling its operation and an established error rate. We offer a relatively simple protocol for such analysis in the hopes that it may provide a base for the eventual development of such a standard. We illustrate the application of our protocol with three case studies from the recent literature.

2. Methodological Overview

These cases involve the early writings of Edgar Allan Poe⁵, the anonymized case of an asylum seeker (cited as "Bilbo Baggins")⁶, and, more famously, the pseudonymous author of *The Cuckoo's Calling*⁷, revealed to be J.K. Rowling of Harry Potter fame. All three cases share several characteristics which may therefore be regarded as "typical"; unlike many literary studies of authorship, these are "verification" problems in which there is really only one candidate author of interest, and therefore available samples. No information is readily available to exclude anyone plausible from authorship (unlike, for example, the Federalist Papers, where scholars readily accepted that authorship was confined to the small group of Hamilton, Madison, and Jay). In each case, the candidate author was an established writer and a baseline of writings by that candidate could be easily obtained and validated.

Previous work has shown^{8 9} that authorship attribution can be performed with relatively high accuracy using a variety of methods. Typical performance on small, closed-class problems is around 80% accuracy.^{10 11} Using ensemble methods such as "mixture-of-experts" can boost performance above the baseline of any individual method. Our proposed protocol, then, is to solve this verification problem by running a number of independent studies as *elimination tests* against an ad-hoc distractor set, to see whether any features set can definitively eliminate the author of interest. Using multiple independent tests provides strong protection both against false acceptance and false rejection errors.

3. Protocol Details

3.1 Ad-hoc distractor set

Most stylometric methods formally choose the most likely author from among a fixed and finite set of candidates based on similarity of writing. While this set is normally chosen based on authors who may actually have had the opportunity to write the disputed document, this is not a formal requirement. From the point of view of stylistic similarity, any two authors or documents can be usefully compared. Koppel et al.¹² noted that randomly chosen authors from the same general field and genre would work as well given repeated measures: "The known text of a snippet's actual author is likely to be the text most similar to the snippet even as we vary the feature set that we use to

represent the texts. Another author's text might happen to be the most similar for one or a few specific feature sets, but it is highly unlikely to be consistently so over many different feature sets." Juola [6] applied the same technique, using newspaper articles scraped from the Web as a baseline against which to compare Baggins' writing.

3.2 Multiple independent elimination tests

The key insight here is that, quoting Koppel, any given wrong author "is highly unlikely to be consistently [similar] over many different feature sets." This insight can be formalized mathematically as follows:

- If a technique is X% accurate, the chance of it being wrong is $(1-X)$. (I.e an 80% chance of being right yields 20% chance of being wrong).
- If two independent techniques are X% accurate, the chance of them both being wrong is $(1-X)^2$.
- If K different techniques are each X% accurate, the chance of them all being wrong is $(1-X)^K$, which becomes arbitrarily small as K increases.

Thus using multiple independent analyses will reduce the chance of false acceptance error to as small a value as desired.

Similarly, false rejection errors can be handled by using a relaxed acceptance criterion, and essentially treating the top few candidates as "successful." This again can be demonstrated rigorously. If our technique is 80% accurate among a set of distractor authors, there is a 20% chance that the most similar author will not be the correct one. But in this case (and with suitable independence assumptions), there will also be an 80% chance that the most similar author among all other authors studied will be the correct one (by assumption), and hence only a 4% chance that the correct author will not be among the top two in the original set. (This chance drops to 0.8% for the top 3.) Thus we can say with high probability that any author not among the top few most similar has been eliminated as a plausible candidate author.

3.3 The proposed protocol formalized

We can thus formalize the proposed authorship analysis protocol as follows: Gather an ad-hoc collection of three to five authors other than the author of interest. Run a number of independent tests of different feature sets to determine which author is most similar to the questioned document on that specific feature. (JGAAP^{13 14} provides a huge number of feature sets from which to choose, and is designed to be extensible to enable people to add additional sets of interest).

Any author not in the two or three most likely candidate authors is eliminated as a potential author. If, after enough experiments have been run, the only author not eliminated is the author of interest, his or her authorship of the questioned documents is deemed confirmed.

3.4 An example (Rowling)

The Galbraith/Rowling case is instructive. In this case, I was provided a distractor set of three authors, all contemporary female British crime writers, so their writings would be comparable to "Galbraith's." Tests were run on four separate feature sets: word lengths, character 4-grams, word pairs, and the 100 most frequent words. Of the four authors, only Rowling was not eliminated by at least one feature set.

We can determine the likelihood of error as follows: Assuming that Rowling was not the author, the probability of her appearing in the top half (top 2 of 4) in any list of candidate authors would be 50%; thus she would have one chance in 16 (approximately 6%) chance of not being eliminated through this procedure.

4. Discussion and Conclusions

Perhaps obviously, there are some caveats to the proposed protocol. The most key is, of course, the implicit assumption of independence. Is it reasonable to believe that the distribution of word lengths is independent of their use of common function words? More importantly, can this belief be validated empirically and justified theoretically? Similarly, there are some numbers in the protocol that may need tightening -- is three to five distractor authors enough? Are five better than three? Can these numbers be justified? We will discuss this further but invite commentary on this point.

It should also be clear that this paper does not *ipso facto* establish a mandatory standard for authorship studies. We invite discussion and even competing proposals, in addition to further studies to establish not only what other protocols might be more accurate, but also which ones are easier to apply, or even more likely to generate useful information (beyond simple authorship). One key aspect of this proposal is that it relies primarily on rank-order statistics and does not take into account the degree of variation; a more sophisticated protocol might use parametric statistics for greater power, at the possible cost of increased complexity.

From a practical standpoint, however, this protocol may represent a substantial maturation of the field. Not only have we used it ourselves, but it has also been used by third parties [5]. The results have been validated by reference to independent ground truths (Rowling acknowledged authorship on July 12, 2013.¹⁵) The results have even been accepted in courts of law. We are thus confident that the proposed protocol will provide a relatively clear-cut way to reduce controversy regarding stylometric authorship attribution and increase its uptake and credibility.

References

1. Kuhn, Thomas S (1996). *The Structure of Scientific Revolutions*. 3rd ed. Chicago, IL: University of Chicago Press.
2. "Thomas Kuhn" (2013). *Stanford Encyclopedia of Philosophy*. plato.stanford.edu/entries/thomas-kuhn/ (Accessed 31 October 2013).
3. Daubert v. Merrell(1993) Dow Pharmaceuticals, 509 U.S. 579
4. Rudman, Joseph (1998). *The state of authorship attribution studies: Some problems and solutions*, Computers and the Humanities, vol. 31, pp. 351–365.
5. Collins, Paul (2013). *Poe's Debut, Hidden in Plain Sight?* New Yorker Blog, 7 October. www.newyorker.com/online/blogs/books/2013/10/edgar-allan-poe-earliest-stories-language-software-investigation.html
6. Juola, Patrick (2013), *Stylometry and Immigration: A Case Study*. Journal of Law & Policy vol 21., pp. 287-298, 2013.
7. Juola, Patrick (2013). *Rowling and 'Galbraith': An Authorial Analysis*. Language Log posting, 16 July 2013. languagelog.ldc.upenn.edu/nll/?p=5315
8. Juola, Patrick (2008), *Authorship Attribution*. Delft: NOW Publishing.
9. Juola, Patrick (2012), *Large-Scale Experiments in Authorship Attribution*. English Studies 93.: 275-283.
10. Juola, Patrick (2012). *An Overview of the Traditional Authorship Attribution Subtask*. Proceedings of PAN 2012, Rome, Italy.
11. Juola, Patrick and Efstathios Stamatatos (2013). *Overview of the Author Identification Task*. PAN/CLEF 2013, Valencia, Spain.
12. Moshe Koppel, Jonathan Schler, Shlomo Argamon & Yaron Winter (2012). *The "Fundamental Problem" of Authorship Attribution*, English Studies, 93:3, 284-291.
13. Juola, Patrick (2009), *JGAAP: A System for Comparative Evaluation of Authorship Attribution*. Proceedings of the Chicago Colloquium on Digital Humanities and Computer Science 1.
14. Juola, Patrick. (2012). *JGAAP 5.3.0: A System for General Purpose Text Classification Experiments*. EACL 2012 Workshop on Computational Approaches to Deception Detection, Avignon, France.

15. Richard Brooks (2013). *Whodunnit? JK Rowling's secret life as wizard crime writer revealed*. Sunday Times article of July 14.

5 Design Rules for Visualizing Text Variant Graphs

Jänicke, Stefan

stjaenicke@informatik.uni-leipzig.de

Image and Signal Processing Group, Institute for Computer Science, Leipzig University, Germany

Geßner, Annette

ageßner@gcdh.de

Göttingen Centre for Digital Humanities, University of Göttingen, Germany

Büchler, Marco

mbuechler@gcdh.de

Göttingen Centre for Digital Humanities, University of Göttingen, Germany

Scheuermann, Gerik

scheuermann@informatik.uni-leipzig.de

Image and Signal Processing Group, Institute for Computer Science, Leipzig University, Germany

1. Motivation

After Schmidt's article ¹ on modelling and representing various versions of text with so called Variant Graphs was published in 2009, web-based tools were developed that utilize and adopt the presented model to facilitate the work with digital editions of text in the browser. CollateX ² is one of these tools. It computes a static, horizontally aligned, directed acyclic graph with vertices showing the various text fragments and edges labeled with edition identifiers connecting subsequent text fragments. The tool Stemmaweb ³ extends the CollateX graph to allow for user-driven annotation and modification of the graph structure (e.g., merging and splitting of vertices). Despite the attached interaction capabilities, it seems that there is little value put on designing the graphs. The purpose of this paper is to raise awareness for improving the readability of Text Variant Graphs. We propose a list of design rules for styling the graph and its vertices and edges to facilitate a rapid comprehension of the underlying alignment structure by the user.

2. Design Rules for Text Variant Graphs

When defining rules for the layout of Text Variant Graphs, we refer to the visualization of CollateX, as its layout is also used in Stemmaweb, and it can be seen as an improvement in comparison to the generated graph of the tool NMerge ⁴ – provided by Schmidt –, where edges carry all types of information.

When we take a look at a resultant CollateX graph (see Figure 1), it is hard to find out how often a text fragment occurs over all given editions. Thus, it is hard to compare, e.g. the numbers of synonyms. The only chance is to count the edition identifiers at the labels of the incoming edges of the desired vertices. But we can easily put this information on the vertex layouts, which leads us to **Rule 1: Vary vertex label sizes!** As Wattenberg proposes for the "Word Tree" ⁵, we suggest weighted vertices also for Text Variant Graphs. The usage of font size as a metaphor to reflect the number of occurrences of individual text fragments helps to immediately differentiate between frequent and infrequent branches of the graph.

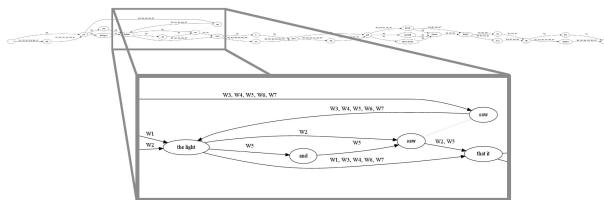


Fig. 1: Fourth Bible verse in 7 different editions: Text Variant Graph computed with CollateX

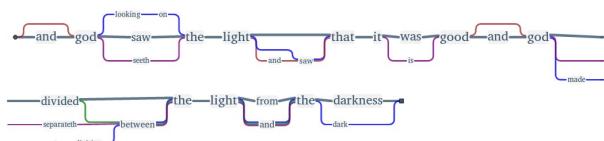


Fig. 2: Fourth Bible verse in 7 different editions: Our visualization

It seems to be obvious to draw the edges of a directed acyclic graph in the shape of arrows. But for what reason, when we know that we read a text with a dedicated writing direction? Then, most probably, the user is supposed to read the graph in the same direction, and it is counterintuitive to move the eyes backwards when reading (in Figure 1, we find an edge from "saw" (right-top) to "the light" (bottom-left) – we call this a backward edge). In graph theory, the common style for a directed acyclic graph is a so called layered graph drawing⁶ with all edges pointing in the same direction. Thus, we define **Rule 2: Abolish backward edges!** When doing so, we can reduce the cognitive load of the visualization by drawing undirected edges instead of arrows.

The labeling of edges with edition identifiers as it is done in CollateX leads to two problems. Firstly, the additional text labels interfere with the vertices' texts. Thus, the reader has to visually separate vertex labels (text fragments) from edge labels (identifiers). Secondly, if lots of editions pass an edge or long edition identifiers are used, the corresponding edge labels become very large. As a consequence, adjacent vertices drift apart and the reader quickly loses the context of a text fragment. Therefore, follows our **Rule 3: Do not label edges!** As an alternative, we suggest drawing an edge for each edition in a different color. However, as the human ability to distinguish colors is limited, it only works well for a small number (<10) of editions. But, with varying stroke styles for edges (e.g., line, dashed line, dotted line) we are again able to increase this number. In any case, a legend is required to map the given styles (or in the CollateX case, the identifiers) to its corresponding edition name.

When analyzing and comparing text editions to each other, the user is often interested in those editions, that deviate from the "general case". Within the Stemmaweb tool⁷, edges, that are passed by most editions, are labeled with "majority", thus, the labels are bundled. When following Rule 3, we receive multiple lines instead of multiple labels, hence, **Rule 4: Bundle major edges!** We highlight both resultant edge types – bundled and unbundled edges – in a different way. Unbundled edges receive saturated colors in visually attracting hues, whereas bundled edges are colored in a plain gray, but drawn with slightly thicker strokes. Thus, the deviations from the general case can be detected easily. When following Rules 3 and 4, we are able to reduce the number of edges to be drawn – and therefore, the cognitive load of our approach – to a passable minimum.

Last but not least, the main problem we identified when reading the horizontal aligned Text Variant Graphs was the required horizontal scrolling. Especially, when the source texts are long, the user quickly loses the context and it is hard to keep track of how individual editions disseminate in the graph. Moreover, a lot of space is wasted since the height of the graph is rather small compared to the height of the screen. The outcome of a survey within the TAdER Project⁸ to avoid horizontal scrolling when reading texts in the browser underpins our hypothesis that the user is accustomed to scroll vertically. Thus, here comes our final **Rule 5: Insert line breaks!** It may

sound tricky to cut the graph into pieces, and thereby, keeping it easily readable. But, why shouldn't we adopt the behavior of a text flowing in a book (with line breaks) for Text Variant Graphs? When following Rule 3, we receive different colored edges (or edge bundles) at the end of each line, so that all paths are visually seizable at the beginning of the next line. This approach supports the user in following individual editions even for large graphs, and the user also receives more context on the screen for a specific position in the graph.

Figures 1 and 2 juxtapose the resultant Text Variant Graphs for the fourth Bible verse with CollateX and our visualization that implements the listed design rules above.

3. Following the Rules: 7 English Translations of the Bible

We are working with seven English translations of the Bible in our project⁹, which turned out to be a very good use-case for the presented visualization approach, not only because the Bible is a very influential and well-known text. Another reason is, that these translations are all derivatives of the same Hebrew and Greek original, often trying hard to preserve the exact wording, and refer to an existing and well structured text, divided into canonical books and verses.

Figure 3 shows a screenshot of the first five Bible verses. After tokenization, normalization and alignment procedures, we layout the resultant directed acyclic graph by following the rules proposed in the previous section. For the seven editions, we chose the following colors of the 12-color palette for categorial usage suggested by Ware¹⁰ to facilitate maximal visual differentiation by the user: red, blue, green, yellow, orange, brown, and purple. To support answering various research questions, the user is able to modify the visualization. Firstly, when the user hovers a vertex all individual edges of the corresponding editions are drawn in the dedicated colors. This mean of interaction helps to highlight the paths containing the dedicated token and to clarify those editions forming majority edges. Secondly, unimportant editions can be removed from the graph. Thirdly, the user is able to select one of the editions as a main branch, so that the corresponding vertices are drawn on the same horizontal level – variations to the other editions can be considered easily.

During the development phase, the humanists of our project steadily evaluated our design, so that the result remains intuitive even for the inexperienced, maybe sceptical user. We strongly recommend such an iterative process when developing visualizations for humanistic applications as it turned out to be very successful. In comparison to a plain graph layout, the presented design for the Text Variant Graph and the project page reminds the user, that it is a book to be read, not just some string of letters – which was a major concern of the humanists.

Our presented approach is still applicable for examples where whole blocks of text have different orders among the various editions, but matching text blocks may strongly drift apart. In the future, we direct our attention on developing algorithms that visually align such structures more properly.

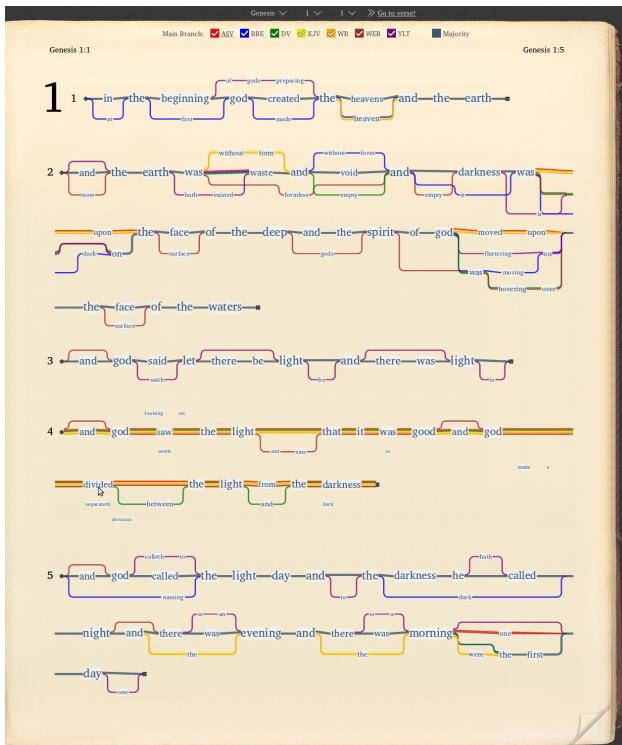


Fig. 3: Screenshot of the Bible use case. In Genesis 1:4, all paths containing the token "divided" are highlighted.

References

1. Desmond Schmidt and Robert Colomb. *A data structure for representing multi-version texts online*. International Journal of Human-Computer Studies, 67(6):497–514, 2009.
2. Ronald H. Dekker and Gregor Middell. *Computer-Supported Collation with CollateX: Managing Textual Variance in an Environment with Varying Requirements*. Supporting Digital Humanities 2011, University of Copenhagen, Denmark, 17–18 November 2011.
3. Tara L. Andrews and Caroline Mac. *Beyond the tree of texts: Building an empirical model of scribal variation through graph analysis of texts and stemmata*. Literary and Linguistic Computing, 2013.
4. Multiversiondocs: Merge, edit and compare N versions in one document. code.google.com/p/multiversiondocs/ (Retrieved 2013-10-30).
5. Martin Wattenberg and Fernanda B. Viégas. The Word Tree, an Interactive Visual Concordance. IEEE Transactions on Visualization and Computer Graphics, 14(6):1221–1228, November 2008.
6. Kozo Sugiyama, Shojiro Tagawa, and Mitsuhiro Toda. *Methods for Visual Understanding of Hierarchical System Structures. Systems, Man and Cybernetics*, IEEE Transactions on, 11(2):109–125, Feb 1981.
7. Stemmaweb – a collection of tools for analysis of collated texts. byzantini.st/stemmaweb/ (Retrieved 2013-10-30).
8. TADER – Text Adaptability is Essential for Reading. www.tader.info/scrolling.html (Retrieved 2013-10-30).
9. Holy Bible – Verses in Various English Translations. informatik.uni-leipzig.de/HolyBible (2013).
10. Desmond Schmidt and Robert Colomb. *A data structure for representing multi-version texts online*. International Journal of Human-Computer Studies, 67(6):497–514, 2009.

A Preparatory Analysis of Peer-Grading for a Digital Humanities MOOC

Kaplan, Frédéric

frederic.kaplan@epfl.ch

EPFL, Switzerland

Bornet, Cyril

cyril.bornet@epfl.ch

EPFL, Switzerland

Introduction

Over the last two years, Massive Open Online Classes (MOOCs) have been unexpectedly successful in convincing large number of students to pursue online courses in a variety of domains. Contrary to the "learn anytime anywhere" moto, this new generation of courses are based on regular assignments that must be completed and corrected on a fixed schedule. Successful courses attracted about 50 000 students in the first week but typically stabilised around 10 000 in the following weeks, as most courses demand significant involvement. With 10 000 students, grading is obviously an issue, and the first successful courses tended to be technical, typically in computer science, where various options for automatic grading system could be envisioned. However, this posed a challenge for humanities courses. The solution that has been investigated for dealing with this issue is peer-grading: having students grade the work of one another. The intuition that this would work was based on some older results showing high correlation between professor grading, peer-grading and self-grading (Wagner et al. 2011¹, Topping 1998²). The generality of this correlation can reasonably be questioned. There is a high chance that peer-grading works for certain domains, or for certain assignment, but not for others. Ideally this should be tested experimentally before launching any large-scale courses.

EPFL is one of the first European schools to experiment with MOOCs in various domains. Since the launch of these first courses, preparing an introductory MOOC on Digital Humanities was one of our top priorities. However, we felt it was important to first validate the kind of peer-grading strategy we were planning to implement on a smaller set of students, to determine if it would actually work for the assignments we envisioned. This motivated the present study which was conducted during the first semester of our masters level introductory course on Digital Humanities at EPFL.

Method

56 students were asked to produce a blogpost in which they had to identify trends in Digital Humanities based on the online proceedings of the DH2013 conference in Nebraska³.

Students had to choose three abstracts from the conference, summarise and compare them, then use the Wordpress blog dh101.ch to publish their post. Students were informed that their post would be graded by the professor but also by other students. Following the usual Swiss grading system, the grade range was from 0 to 6. The students were informed that only the grade given by the professor would count for their semester results but that 10% of their semester results depended on whether they took the peer-reviewing seriously.

The grading criteria was presented in detail to the class at the same time. Students had to check whether the blog post followed the guidelines of the assignment (discussing three articles, identifying a trend) (4 points); whether the English was correct and clearly understandable (+0.5); whether the keywords and post layout were adapted to its content (+0.5); whether the post was not just a summary of the three articles but really compared them, and, more subjectively, whether the post's content was well discussed (+0.5) and the identified trend interesting (+0.5). The students were also asked to verify that the blog post did not contained plagiarised content.

Each student had to anonymously grade five randomly chosen blog posts. In order to simplify the task and to reduce the risk of manipulation errors, we developed a simple dedicated Wordpress app⁴ to organise this process. Students used their Wordpress account to log in and by doing so accessed a page listing the five posts that were assigned to

them, as well as one checkbox per criteria to be checked. This assignment process was done beforehand in an automated way: each paper was assigned once to the professor and randomly to five students, but under the constraint that no student could get their own paper and no more than five papers in all. Although we are aware that more sophisticated systems exist for assigning work in peer-reviewing processes (e.g. The Caesar system developed at MIT⁵), we assumed that this random assignment process was relevant in this context, given the uniform nature of the content to be graded. The professor graded all the blog posts without any information on the results of the peer-grading process.

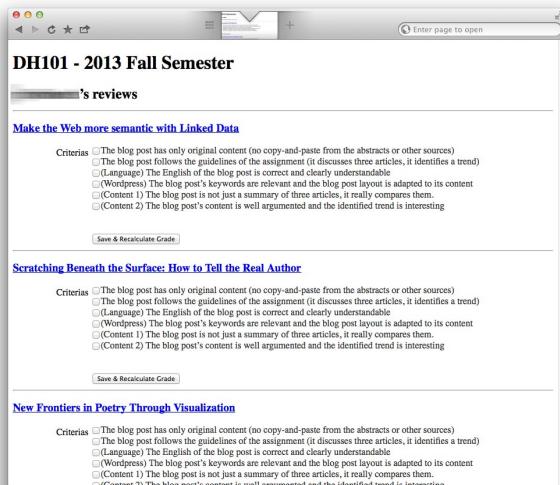


Fig. 1: Wordpress app for peer-grading

Results and Discussion

52 blogs posts were produced and published on the public website dh101.ch. 47 students completed the peer review grading, and all of them graded all of their five papers. Three students accessed the website without grading any papers, and two didn't even attempt to log in.

The process drew a lot of interest and questions from the students. The criteria of the grading grid were questioned by many students, in particular the one linked with the more subjective evaluation of the posts (several students thought this was unfair). Other studies have shown that students' writing and understanding in core courses can be improved through peer ranking (Rhett A. et al 2005⁶). Although it is difficult to measure, it seems that the precise explication of the grading criteria imposed by the peer-review system had a positive effect on the quality on the posts when compared to the last year's course, when this system was not included.

We measured a strong correlation between the average of the peer-graded marks and the mark of the professor ($r(50) = 0.39$, $p < .01$).

Figure 2 shows the level of matching between the professor's grade and the peer graded marks after normalisation to the closest half point. In 38% of the case, the peer-graded mark is the same as the professor's, in another 38% of the cases the mark shows a 0.5 difference. The remaining 24% of cases show a larger difference. These latter cases mostly correspond to situations where either the professor or one of the students concluded that the post did not respect the instructions for the exercise and therefore gave a sanctioning mark (0 in case of plagiarism, 3 in the case of uncompleted post).

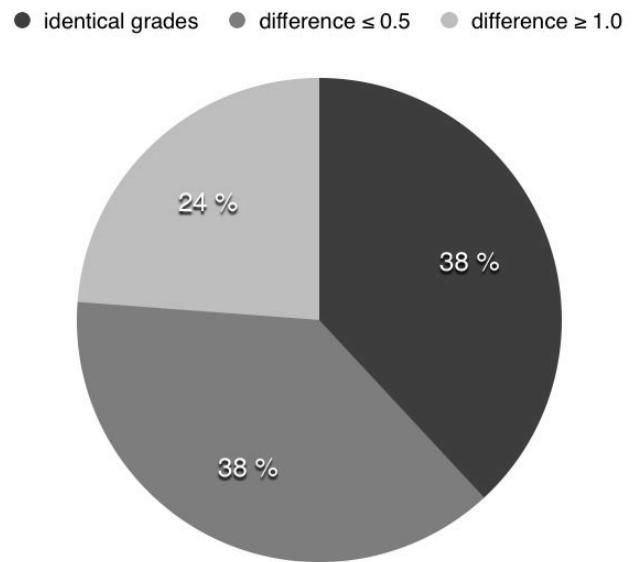


Fig. 2: Level of matching between professor and student grades

This relationship can also be shown by breaking down the ratios of validated criteria:

Figure 3 shows the average grades given by students (sorted ascending), the squared mean error intervals for each of them and the professor's gradings (x marks), confirming that 76% of the papers received an average grading within 0.5 to the professor's.

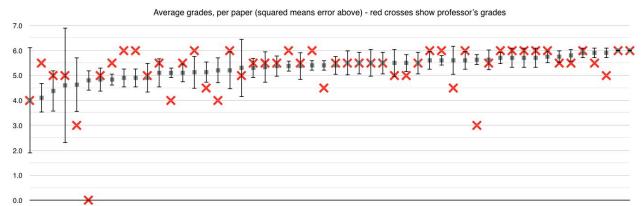


Fig. 3: Average grades, per paper

Figure 4 shows a comparison between the grading criteria used by the students and the professor's grading. The distribution is visually similar (we cannot perform a more detailed statistical analysis because of the sample size).

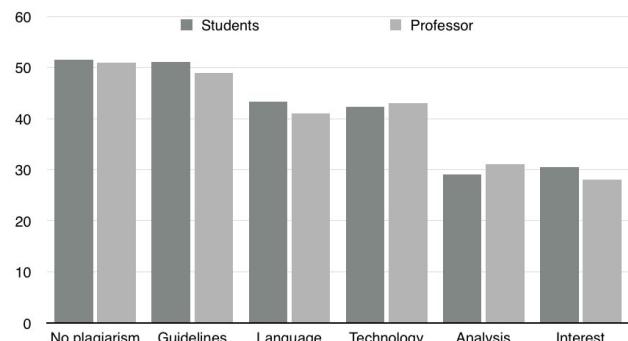


Fig. 4: Criterias Distribution

Figure 5 presents the correlation between the student's grade and their own grading. Although it is not obvious at first sight, there is a marginal significant negative correlation between these two variable datasets ($r(50) = -0.26$, $p = .07$). This could suggest that students who wrote good papers are more critical of their peers.

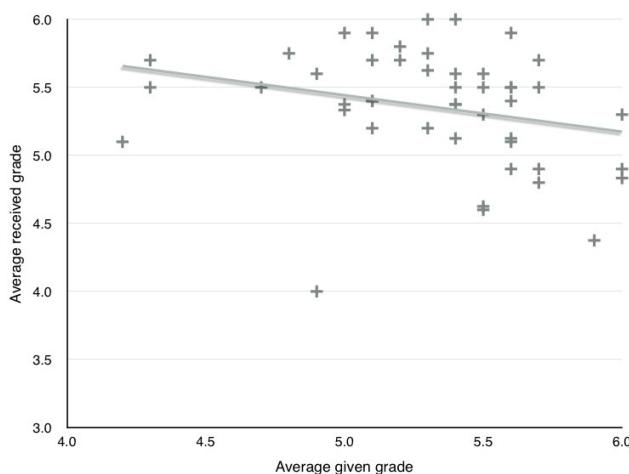


Fig. 5: Correlation between received and given grades

Figure 6 presents the grades in the order published (the first posts are the ones submitted the earliest by the students). The professor graded the posts following this order. As expected, no correlation exists between the order and the grades in the students' gradings, as they were assigned randomly. However, there seems to be a tendency towards lower grades in the professor's grading sequence. This could be explained if a correlation existed between the quality of the post and their publication time (the best students would publish the first). However, this correlation was not found in the student's grading. This could suggest a potential temporal bias observed through the fact that the last evaluations were tendentially lower than the first ones (the professor becoming more critical).

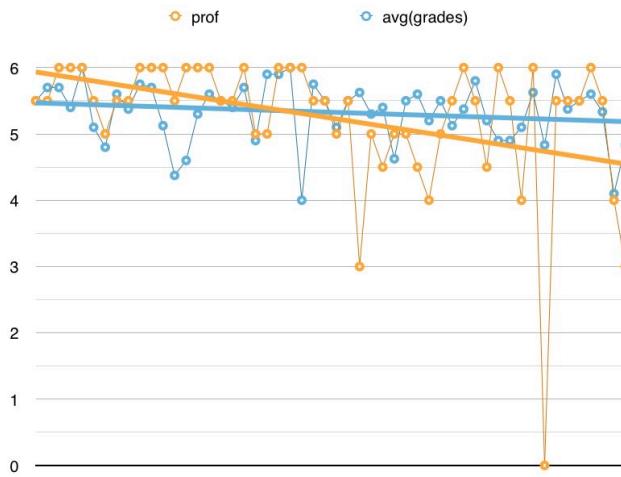


Fig. 6: Evolution of grades with time

Conclusion

This article presents a preliminary study for a Digital Humanities MOOC conducted on a class of 56 students. In the present context, we can conclude that peer-grading is highly correlated to the professor's grading. In about $\frac{1}{4}$ of the cases, the grades obtained by this method are within a range of 0.5 points of the professor's grading. Qualitative observations tend to show that the quality of the posts increased when compared to the previous year, likely because students had to reflect on the grading criteria and were cautious of producing good work when this work would be evaluated by their peers. In addition, our study may suggest some possible temporal biases in the way the professor grades a long sequence of work, reinforcing the idea that peer-grading may not only be an interesting positive alternative to traditional grading but also that it may, in some cases, be less biased.

Nevertheless, these preliminary results, dependent of the particular context of this study, should be extrapolated with care and would not eliminate the need to conduct regular quality evaluation in the context of a MOOC on Digital Humanities. Indeed, these results do not guarantee that the same peer grading method could scale to a 10 000 student MOOCs without problem. As the number of students increases, and the cultural backgrounds and linguistic competencies diversify, part of the behavioural homogeneity that we observed in this pre-study may no longer be valid. For this reason, this research should definitely be completed with a *a posteriori* study testing the efficiency of peer-grading with a similar method in a randomly chosen set of learners of the entire MOOC.

Acknowledgments:

The authors would like to thank Andrea Mazzei for fruitful discussions about the analysis of the results of this preliminary study.

References

1. Wagner, M.L., Churl Suh, D., Cruz, S. (2011). Peer- and Self-Grading Compared to Faculty Grading. American Journal of Pharmaceutical Education 75, 130.
2. Topping, K. (1998). Peer Assessment between Students in Colleges and Universities. Review of Educational Research 68, 249–276.
3. DH2013 Abstracts, dh2013.unl.edu/abstracts/
4. Wordpress Login API, codex.wordpress.org/Plugin_API/Action_Reference/wp_login
5. Tang, M. (2011), Caesar: A social code review tool for programming education, Master project, MIT
6. Rhett A. et al (2005), Using peer ranking to enhance student writing, IOP Science

WÆΓNING: A Conceptual Parsing of ASCII Character Substitutions

Katelnikoff, Joel

University of Alberta, Canada

This presentation will investigate K-Rad character substitutions as they were utilized in serialized ASCII text publications during the 1980s and 1990s. These text files, influenced by a genealogy of informational manuals on the topics of hacking and phreaking, sought to take writing beyond transparent signification, imposing hackerly techniques on to the text itself, by extending, disrupting, and hypermediating codes of discourse. In conventional informational writing, disruption and hypermediation are considered to impede signification; in K-Rad texts, they become key signifying agents.

K-Rad writing attempts to push boundaries, to present the unrepresentable, and to break all codes, whether legal, moral, linguistic, or typographic. Although the style finds precursors in the underground discourse of piracy and software cracking, K-Rad was not universally esteemed in these contexts. For example, as Rabid Rasta says in "The Real Pirate's Guide" (1984):

Real Pirates Don't Say "K-K00L", "K- Awesome", "X10Der", "L8R0N", Or Anything Of The Sort.

Real Pirates Know The Difference Between "F" And "Ph" (I.E."Philes", "Phuck", "Fone", Etc.).

In Rabid Rasta's opinion, unconventional spelling primarily demonstrates a pirate to not be "REAL." His attitude attempts to legalize the spelling of words, which, according to Roland Barthes, "keeps the scribe from enjoying writing, that euphoric gesture which permits putting into the tracing of a word *a little more than its mere intention to communicate*" ("Freedom to Write" 45). Such substitutions, particularly in published ASCII text files, are rarely accidental; these substitutions extend beyond conventional meaning, into the realm of the specifically technologized word. Consider, for example, the following example from "tHe PHiRzT StEp!!" by (V)[](_>|#|_) (1994):

iFh i HAvEN'T g0T y0U t0 S+oPA rEA|>iN' YeT tHeN i
FiNK 0NiY dA Pe0PIE >tHAt wErE ELiT E eNUFF t0 tA</3 iT
sTaYeeD. On one hand, it is hard for the human eye to read, because it corrupts the alphabet, adding unfamiliar characters to the familiar ones, repurposing characters, using them in new contexts and infusing them with new meanings so as to make them strange. On the other hand, in an era of command line instruction, where computers were incapable of recognizing these visually corrupted renderings, it was *only* the human eye that could parse them. These words, which we cannot simply reduce to their legalized spellings, demand that we decipher the text on a character-by-character basis.

If the writing seems to falter at the sentence level, this is only because the writing style focuses primarily on the grapheme and the word, constituting an extreme close-up that blurs the shape of the sentence, the paragraph, and the text as a whole. Although we often, in the tradition of Saussure, consider the word to be the smallest unit of meaning, K-Rad graphemes become narrative elements, turning each word into a story. In this context, graphemes take on value within the world of the word. Stanley Fish says, "A reader's response to the fifth word in a line or sentence is to a large extent the product of his responses to words one, two, three, and four" (*Is There a Text in This Class?* 27). With K-Rad orthography, we become aware that this process also takes place at the level of the grapheme—a reader's response to the fifth grapheme is determined largely by graphemes one, two, three, and four. Every word requires active contextualization, active decipherment, a flow that does not merely move forward from left to right across each line, but scans in multiple directions within each assemblage of characters.

To the uninitiated, the writing might present itself only as a kind of line noise. Even for a well-versed reader of ASCII, the reading process never becomes a purely linear one in which words travel left to right without impediment, but, as Viktor Shklovsky says, "Art is not a march set to music, but rather a walking dance to be experienced or, more accurately, a movement of the body, whose very essence it is to be experienced through the senses" (*Theory of Prose* 22). Although texts might be said to transmit information or communicate, we cannot say that this is *all that they do*, unless we feel comfortable to ignore all of the text's non-informational signifying elements. As Jerome McGann says:

"When we imagine texts as transmitters we are not wrong in our imagination, but we are narrow—and much narrower than we should be if we wish to understand how texts work. Indeed, we easily confuse investigations of textuality when we study texts as machines for carrying messages. In the reading of poetry—those paradigm texts—this kind of confusion typically arises in thematic studies, where the "meaning(s)" of the texts are pursued. In poems, however, "meaning" is mistakenly conceived if it is conceived as "message." Rather, "meaning" in poetry is part of the poetical medium; it is a textual feature, like the work's phonetic patterns, or like its various visual components. (*The Textual Condition* 14-15)"

The ASCII text file itself, as a medium, signifies a faith in the new technology of telecommunication and advances ASCII as a preferred means of communication. These texts are not only about their linguistic messages, but also about *acquiring, arranging, and moving text*, activities that constitute messages in themselves. And beyond all of this, we might also consider K-Rad texts as concrete poetry, a kind of ambient writing, a kind of visual arrangement that can make the initiate reader feel that they are beholding computer code itself and gazing upon some hackerly art. In a 1995 issue of *Maclean's* magazine, Joe Chidley described the actions of the hacker:

"His fingers trip lightly over the keyboard. With the punch of a return key, a string of characters – writ in the arcane language of computers – scrolls onto the black- and-white display in front of him. "OK," he says, "I'm in." Suddenly, horizontal rows of letters and numbers scroll from left to right across the screen – meaningless to the uninitiated eye. But for the hacker, the mishmash of data contains seductive, perhaps lucrative secrets. ("Cracking the Net" 54)"

Lucrative secrets might be found in the technology of word processing, or the American Standard Code for Information Interchange, or the keyboard, or the technology of writing,

or the technology of language in general. K-Rad adds extra texture to the word, complicating meaning and denaturalizing the basic elements of how we communicate. This kind of writing is radical not merely because of *what it says* but *how it says*. At the 2014 Digital Humanities Conference, I will present a series of K-Rad ASCII text files, demonstrating a sequence of increasingly baroque substitutions and suggesting conceptual reading practices that we might use to engage with these early avant-garde digital texts and also with conventional literature. These stylized visions do not only affect our reading practices here, but our reading practices everywhere, even in when dealing with texts that are not self-consciously stylized. These highly-coded and hypermediated texts confront us with the fact that there is no such thing as neutral writing, no such thing as neutral discourse, and once we have seen the word in its codified and disruptive form, there is no way to return to our previous unconscious state.

This topic of this presentation is based on "HACK," a chapter from my recent dissertation, *SCROLL / NETWORK / HACK: A Poetics of ASCII Literature* (1983-1989).

Problems in Encoding Documents of Early Modern Japanese

Kawase, Akihiro

a_kawase@nijal.ac.jp

National Institute for Japanese Language and Linguistics, Japan

Ichimura, Taro

tichimura@nijal.ac.jp

National Institute for Japanese Language and Linguistics, Japan

Ogiso, Toshinobu

toogiso@nijal.ac.jp

National Institute for Japanese Language and Linguistics, Japan

1. Introduction

As part of its KOTONOHA (meaning 'words of the language' in classical Japanese) Project, the National Institute of Japanese Language and Linguistics (NINJAL) is conducting a morphological analysis of Japanese classics. Both the selection and digitization of classic documents are required in order to execute the analysis correctly. Digitization and morphological analysis have been done thus far on the literature of several ages and styles (Ogiso et al., 2012; Ichimura et al., 2012; 2013), and various text corpora have been published. These text corpora are marked up with NINJAL's original Document Type Definition (DTD). However, some elements are used in common on all corpora but basically are not unified nor standardized. Under this circumstance, it enables structural analysis and string extraction from a single corpus but causes problems with structural comparison and numerical analyses between several corpora. Thus, it is necessary to design and mark up a unified definition from a higher level in order to conduct analyses concurrently.

In the previous study, we examined the possibility of converting classic documents with TEI-compliant XML (Kawase et al., 2013), in particular, by designing a tag-set and strictly structuring an old wood-block printed book from Sharebon's "*Keisei-kai Futasuji-no-michi*" (published in 1798, e.g. Fig. 1) as a model case, and sorted out this problem. This work aims to provide further insights into those problems and propose a solution.

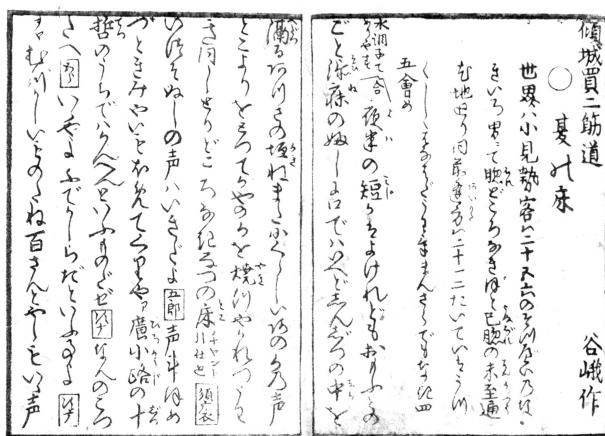


Fig. 1: Excerpt image of Sharebon's "Keisei-kai Futasuji-michi" (owned by NINJAL)

2. Significance of Encoding Classic Documents

The reason for choosing *Sharebon* as a subject is because it is important primary linguistic data from the early modern period (Keene, 1999) and it has the following three features: (1) published in the broad age from the 18th to the first half of the 19th century; (2) uses colloquial words and expressions among conversations between characters; (3) abundantly describes things in Edo (former name of Tokyo) language and Kamigata (Kyoto-Osaka region) language.

Therefore, describing *Sharebon* in a machine-readable manner has profound significance not only for bringing numerical analysis of unexplained colloquial expressions of the modern period into reality but also for facilitating humanities research in language history and descriptive bibliography. Furthermore, since many texts that have a similar structure to *Sharebon* were published over the same period of time, this study will offer a common format for archiving coeval literature.

3. Issues in Encoding *Sharebon*

In order to analyze the manuscript from a corpus linguistic viewpoint, not only outward mark-ups but also internal mark-ups of both document structure and linguistic structure are required. In general, *Sharebon* is composed of a combination of three parts: the front matter; the narrative body that contains colloquial expressions and descriptive texts; and the back matter. For instance, in the case of "Keisei-kai Futasuji-michi", the narrative body is made up of two chapters, 'Natsu-no-toko' (Summer Alcove) and 'Fuyu-no-toko' (Winter Alcove). Since the composition of this whole document is well accorded with the composition of an orthodox Western manuscript, we may mark up most of the elements in reference to TEI P5: Guidelines (Burnard and Bauman, 2007). Table 1 shows the list of elements with an explication of their roles used to mark up the document.

The problems incurred in the process of encoding *Sharebon* can be broadly classified into two matters: (A) text formatting on a #PCDATA (character data) level; and (B) structuring ruby annotations. We will discuss these two matters in more detail in the following sections.

Element	Role	Element	Role
<text>	whole text	<s>	sentence unit
<front>	front matter	<l>	verse line
<body>	body part	<rs>	referencing string
<back>	back matter	<w>	ruby annotation*
<div>	divisions	<g>	extended character
<head>	title	<pb>	page break
<p>	paragraph	<cb>	column break
<q>	quote	<lb>	line break
<stage>	interlinear notes	#PCDATA	character data

Fig. 2: List of elements with an explication of their roles to mark up Sharebon

3.1 Text Formatting on #PCDATA

To design and assure a high-quality corpus in terms of linguistic resources, it is necessary to index information about lexical morphemes (e.g. word class, inflected forms, pronunciations) at the body-text level. However, since there are three problems in the sentences of early modern Japanese, it is difficult to conduct morphological analysis properly: (A-1) voice markings called *dakuten* are missing where they should be; (A-2) corresponding phonetic characters called *hiragana* and *katakana* are not unified; (A-3) iteration symbols called *odoriji* are unmodified.

For example, *dakuten* makes a phonetic shift from *ka*, *ki*, *ku*, *ke*, *ko* (written in hiragana letters as か, き, く, け, こ) to *ga*, *gi*, *gu*, *ge*, *go* (rendered as か~, き~, く~, け~, こ~).

3.2 Structuring Ruby Annotations

Such annotations are the small-text characters rendered alongside the base text. In vertical writing, ruby is typically printed on the right-hand side. However, since the three problems exist, it is difficult to mark up both outward information and the linguistic structure simultaneously: (B-1) ruby text carries extended characters (*gaiji*), damaged and missing characters, and some typographical errors and omissions; (B-2) words of such text and base text are not in one-to-one correspondence; (B-3) There can also be another ruby annotation on the left-hand side simultaneously.

4. Means for Solving the Problem

4.1 Text Formatting on #PCDATA

At NINJAL, for problems (A-1), (A-2), and (A-3), the character level is corrected using the original elements of <vMark>, <kana>, and <odoriji>, respectively (Ichimura et al., 2012; 2013). We accurately describe the text formatting in a uniform way by combining the TEI elements <seg> and <choice> (e.g. Fig. 2).

According to TEI P5: Guidelines (Burnard and Bauman, 2007), <seg> (arbitrary segment) represents any segmentation of text below the 'chunk' level, and <choice> groups a number of alternative encodings at the same point in the text. Distinctions in the corrections of (A-1), (A-2), and (A-3) can be shown by writing @type, holding selectable values from vMark, kana, and odoriji, to the attribute of element <seg>. The original text marked with TEI element <orig> (original form), along with the corrected text (the optimal version for morphological analysis) marked with TEI element <reg> (regularization) are written below the <choice> level.

This description policy preserves both the outward and linguistic structure, even if problems (A-1), (A-2), and (A-3) carry over extended characters or omissions in the original manuscript, and brings morphological analysis into reality as well.

4.2 Structuring Ruby Annotations

At NINJAL, regarding structuring ruby annotation, since priority is given to morphological analysis, words and characters are encoded using the original defined element <ruby> with an attribute @rubyText (e.g. Fig. 3a). To express this structure with TEI-compliant XML, we can simply substitute the element <ruby> with <w> (word) and the attribute @rubyText with @ana (e.g. Fig. 3b). However, the important ruby text information is described inside @ana as a value, we cannot solve problem (B-1) under this policy. Depending on the technology, CSS, XHTML, and HTML5 platforms might be considered as

alternatives in order to structure ruby annotation (Benoit, 2010) (e.g. Fig. 3c).

Here, each `<ruby>`, `<rt>`, and `<rp>` represent inline elements that contain base text with ruby annotation, the superscript which comes over the base text, and ruby parentheses which are used to wrap around opening and closing parentheses `<rt>`, respectively. However, since the ruby text comes over the #PCDATA level, we cannot solve problem (B-2) under this policy. In addition, we have to solve the above problem and problem (B-3), securing the coexistence of ruby on the right-hand side and left-hand side, simultaneously.

<pre><seg type="vMark"> <choice> <orig>か</orig> <reg>か</reg> </choice> </seg></pre>	<pre><seg type="kana"> <choice> <orig>か</orig> <reg>カ</reg> </choice> </seg></pre>	<pre><seg type="odoriji"> <choice> <orig> ></orig> <reg>か</reg> </choice> </seg></pre>
(A-1)	(A-2)	(A-3)

Fig. 3: Examples for encoding ruby annotations

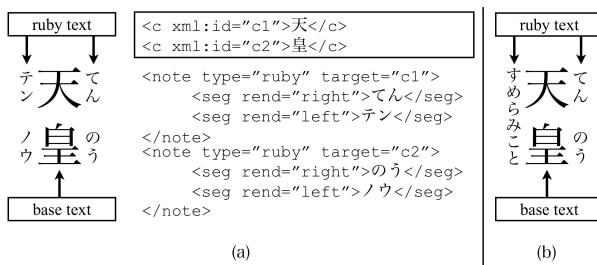


Fig. 4: Examples for ruby annotations on both sides

Imaginably, as shown in Fig. 4a, we will be able to suggest one solution to express the character on base text in a stand-off fashion, in case of the base text carries double ruby on the both sides. However, the above suggestion may work only the character string and both sides of ruby correspond each other. As shown in Fig. 4b, when ruby on either side does not fit or exceed the each target character, this suggestion would be inapplicable.

Currently, we have difficulty in choosing a path that allows specifying both the object outline and structure of Japanese language together; therefore, a new definition over ruby annotation is needed as a means to this end. As shown in Fig. 5, our current solution is to mark up the manuscript in a way which makes it possible to output two types of XML file that specify the information of the object and the structure of Japanese language separately, rather than to fulfill both at the same time.

element of <seg> and <choice>, was considered. For problem (B), major difficulties of structuring both outward information and the linguistic structure of Japanese documents at the same time were presented, and problems to solve were pointed out. Especially, ruby annotation is absolutely imperative for Japanese documents. This extra-textual addition is common and employed almost everywhere from historical documents to modern comics for highlighting the pronunciation or meaning of a word. Our future challenge is to examine a compromise that satisfies both structures simultaneously while working out the problems with the TEI Council.

Funding

This work was supported by the collaborative research project “Design of a Diachronic Corpus” and “Study of the History of the Japanese Language Using Statistics and Machine Learning” carried at the National Institute for Japanese Language and Linguistics.

References

- Benoit, G.** (2010). *Expanding, Facilitating, and Applying Ruby to Explore User Engagement with Encoded Texts*, <http://web.simmons.edu/~benoit/rc/ruby/Ruby-Poster.pdf> (accessed 15 October 2013).

Burnard, L. and Bauman, S.(2007).TEI P5: *Guidelines for Electronic Text Encoding and Interchange*, Text Encoding Initiative.Arlington, MA: TEI Consortium. <http://www.tei-c.org/Guidelines/P5/> (accessed 15 October 2013).

Ichimura, T., Kawase, A. and Ogiso, T. (2012). *Structuring Colloquial Early Modern Japanese Text and Its Issues of Definition*, Proceedings of the 96th IPSJ SIG Computers and the Humanities, CH96: 1-8.

Ichimura, T., Kawase, A. and Ogiso, T. (2013). *Structuring the Corpus of Share-bon*, Proceedings of the 3rd Workshop on Corpus Japanese Linguistics, 249-258.

Kawase, A., Ichimura, T. and Ogiso, T. (2013). *Problems in TEI P5 Encoding on Colloquial Japanese Documents of the Early Modern Period*, Proceedings of the IPSJ SIG-CH/PNC/ECAI/CIAS Joint Symposium 2013, 7-12.

Keene, D. (1999). *A History of Japanese Literature Volume 2: World Within Walls*, Columbia University Press, New York.

Ogiso, T., Komachi, M., Den, Y. and Matsumoto, Y. (2012). *UniDic for Early Middle Japanese*, Proceedings of the 8th International Conference on Language Resource and Evaluation (LREC), 911-915

History All Around Us: Towards Best Practices for Augmented Reality for Public History and Cultural Empowerment

Kee, Kevin Bradley
Brock University, Canada

Compeau, Timothy
Western University. Canada

Poitras, Eric
McGill University, Canada

The desire to reveal the history all around us, to see into the past, is as old as civilization.

With the emergence of augmented reality (AR) technology, historians and public humanities professionals are exploring new ways to research, teach and learn about the past. AR applications augment the physical world by embedding it with digital data, networking, communication abilities and enhanced properties (Mackay 1996). When harnessed for history and public humanities, AR represents a disruptive way of accessing

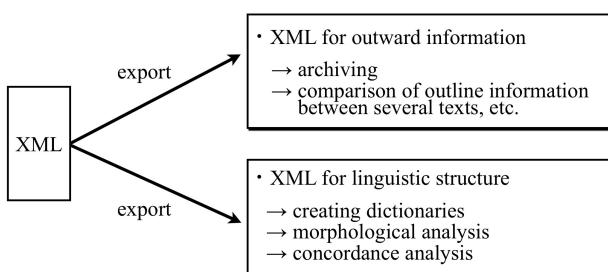


Fig. 5: Solution for problem (B): exporting two types of XML files

5. Conclusion

In this study, developing a corpus for historical documents in a comprehensive and versatile way was considered the ultimate goal. We devised a tag-set based on the TEI-element, and structured *Sharebon's "Keisei-kai Futasuji-michi"* as a model case. We examined the problems of (A) text formatting on #PCDATA level and (B) structuring ruby annotations which were previously unsolved. For problem (A), a concrete policy to bring morphological analysis into reality, by combing the

knowledge, making discoveries, and communicating history in new and imaginative ways. In "History All Around Us: Towards Best Practices for Augmented Reality for Public History and Cultural Empowerment", the authors reflect on the design, development and testing of two location-based AR applications, and propose best practices for using AR to enrich our understanding of history, and support the cultural empowerment of citizens.

The paper is organized in three parts. In part one, the authors draw on the digital humanities to form conclusions about best practices for AR design and development with a focus on two iPhone applications (see www.historytours.com). These apps introduce visitors to the history of the villages of Queenston and Niagara-on-the-Lake, Canada (the latter of which hosts more than 2 million visitors a year). The authors reflect on the effectiveness of the design of these apps, and the development team's decision to offer visitors to these villages two kinds of experiences. The first experience, called "Roam Mode", follows the user's directions. It functions like an on-demand tour guide, providing the user with information about the historical buildings and objects that surround her. Apps that employ strategies similar to "Roam Mode" are now ubiquitous. Much less common is the second experience, called "Quest Mode". It uses gamification (Deterding et al, 2011) to draw the user into exploring the history of the villages to solve long-standing mysteries. Inspired by real-life events, "Quest Mode" features historical personages linked by sometimes real, and sometimes imaginary events. In the case of Queenston, for instance, the user is enlisted to help solve the mystery of who bombed the nearby monument to a local war hero. Along the route, the user must solve puzzles; for example, the user must spot the differences between the real Fort Niagara, which stands watch imposingly from the banks of the United States, to an image of the fort that has been "discovered" by local historians.

The authors outline several best practices for augmented reality design and development for public history, giving special attention to "Quest Mode" and the concept of gamification. They note that the roots of gamification theory lie in the work of game theorists who have separated virtual environment games from real-world environments. These scholars have spent a decade attempting to define what games are, and therefore what they are not (Juul, 2005; Zimmerman, 2004; Pearce, 2004), and have pointed out the many ways that space (virtual, or real world) is treated differently (McGregor, 2006). The authors disagree with a strict adherence to this virtual-/real-world distinction, and suggest that there is much to learn by considering virtual and real-world environments together, because best practices for the development of experiences in one can be applied to the other, resulting in a better understanding of the attributes of gamified augmented reality applications, and contributing to the emergence of (and our understanding of) new forms of expression.

At the same time, the authors urge caution about drawing too heavily on established game theory, and making unwarranted connections between games in virtual space and (augmented) real space. Foundational authors in game studies, such as Roger Callois (1961) and Johan Huizinga (1964) have classified games as activities essentially separate from normal life. But gamified augmented reality applications, such as the two iPhone applications addressed in this paper, often take place during working hours on city streets, and involve "players" playing in the midst of everyday life. (de Souza e Silva, 2008). Similarly, scholars have classified games as primarily escapist – causing players to disengage from their "real world" communities – but the application of interactivity to political and social issues has shown the potential of gamified augmented reality environments for collective cultural empowerment.

In part two of the paper, the authors draw on social science theory and methodology to provide a preliminary report on the testing of these iPhone apps. The authors note that while development of these kinds of digital environments (including, but not limited to augmented reality applications) is now commonplace in the digital humanities, rigorous testing for user engagement and learning within these applications is less common. They note that there is little research that addresses the assessment and appraisal of learning and engagement

in augmented reality applications on mobile platforms, and highlight the need for principled and replicable methodologies.

The authors report on their progress towards evaluating the two augmented reality iPhone applications in terms of fostering learning and engagement. Specifically, the authors report on their use, for the purposes of testing, of two theoretical frameworks: i. the Benchmarks of Historical Thinking as outlined by Peter Seixas (2004, 2011; Peck & Seixas, 2008) and; ii. the Control-Value Theory of Emotions as described by Reinhard Pekrun (2009). In addition, the authors outline how they intend to promote learning and engagement in these kinds of apps by embedding dynamic assessment mechanisms to adaptively modify the content that is provided to the users, and improve user experience.

In part three, the authors briefly speculate about the ways in which the imminent (2014) arrival of commercial augmented reality platforms such as Google Glass (<http://www.google.com/glass/start/>), and connected Google Glass heritage- and history- themed applications such as "Field Trip" (<https://play.google.com/store/apps/details?id=com.nianticproject.scout&hl=en>) will transform the ways in which digital historians and digital humanists develop and use ubiquitous computing and augmented reality for cultural empowerment.

References

- Callois, R.** (1961). *Man, play and games*. (Meyer Barash, Trans.). New York: Free Press of Glencoe.
- Deterding, S., Dixon, D., Khaled, R. and Nacke, L.** (2011). "From game design elements to gamefulness: Defining "gamification"". Proceedings of the 15th International Academic MindTrek Conference.
- Huizinga, J.** (1964). *Homo ludens: a study of the play-element in culture*. (Translated from the German). Boston: Beacon Press.
- Juul, J.** (2005). *Half-real: video games between real rules and fictional worlds*. Cambridge, MA: MIT Press.
- Mackay, W.** (1996). "Augmenting reality: A new paradigm for interacting with computers". La Recherche (March).
- McGregor, G.L.** (2006). "Architecture, space and gameplay in world of warcraft and battle for middle earth". Proceedings of the 2006 International Conference on Game Research and Development, Perth, Australia. Retrieved October 31, 2013 from <http://www.users.on.net/~georgia88/files/Architecture,%20Space%20and%20Gameplay%20-%20Georgia%20Leigh%20McGregor.pdf>
- Pearce, C.** (2004). "Towards a theory of game / responses". In Noah Wardrip-Fruin and Pat Harrigan, (Eds.). First Person: New Media as Story, Performance, and Game Cambridge: MIT Press.
- Peck, C., and Seixas, P.** (2008). *Benchmarks of historical thinking: First steps*. Canadian Journal of Education, 31:4.
- Pekrun, R., and Stephens, E. J.** (2009). *Goals, emotions, and emotion regulation: Perspectives of the control-value theory of achievement emotions*. Human Development, 52, 357-365.
- Seixas, P.** (Ed.). (2004). *Theorizing Historical Consciousness*. Toronto: University of Toronto Press
- Seixas, P.** (2011). "Assessment of historical thinking". In Penney Clark (Ed.), *New Possibilities for the Past* (pp. 139-153). Vancouver: UBC Press.
- de Souza e Silva, A.** (2008). "Hybrid reality and location-based gaming: Redefining mobility and game spaces in urban environments". *Simulation & Gaming* 39:1 (March).
- Zimmerman, E.** (2004). "Narrative, interactivity, play, and games". In Noah Wardrip- Fruin and Pat Harrigan, (Eds.), First Person: New Media as Story, Performance, and Game Cambridge: MIT Press.

Aiding Modern Textual Scholarship using a Virtual Hinman Collator

Kejriwal, Gaurav

Texas A&M University, United States of America

Furuta, Richard

Texas A&M University, United States of America

Olivieri, Ryan

Texas A&M University, United States of America

Introduction

Collation is an important step in textual criticism and it is most often an arduous task for most scholars involved in scholarly edition. Unsworth includes "collation" as one of the scholarly primitives which have been basic to scholarship across eras and media. Textual variation has been a pervasive problem affecting literary text since the invention of writing. It can arise in two forms - either due to repeated copying of a manuscript such as the variants in the First Folio of Shakespeare or those inadvertently inserted by the author/copyist such as the changes made in Mary Shelley's Frankenstein. In the first case collation aids the scholar in generating a critical edition. In the latter case, collation can help the scholar understand the author's purpose. Finding variations is important for research in bibliography and book history as well.

Most of the focus in digital humanities until now has been on making documents available digitally. Much less focus has been on actually supporting the process of scholarly research. The area of collation too awaits a lot more from technology. In the late 1940s Charlton Hinman invented the Hinman collator. Using optical means, it allowed manual comparison of separate copies of a text in order to detect differences. Descendants of the Hinman, collators like the Mcleod collator, the Lindstrand collator and the Hailey's Comet, are still used today. Mechanical collators take time to setup correctly, cannot be used on varying fonts, can damage the books, and are expensive and not very portable. Another approach is to perform collation using tools like Juxta or Collate X on text obtained by transcription or by OCR. This method is flawed by the limits of OCR technology and human error. The Sapheos project approaches the collation problem interestingly by attempting to unwarped the pages and registering them using SIFT key points, but this approach will fail if the text differs significantly.

Most of the tools used today are standalone which inhibits collaboration. Also scholars prefer original copies or facsimiles instead of OCR or transcription versions and most of these tools don't support that.

This work proposes to address these problems. A prototype of the virtual Hinman (vHinman) collator was created and user-evaluation was conducted amongst scholars experienced with collation work. Image-matching algorithms along with context information are used to match words and the tool was integrated into the creativity support environment CritSpace. Moreover, CritSpace provides the functionality to easily extend the tool to support collating multiple (>2) copies.

Methodology

We developed and evaluated some approaches towards comparing page-images.

Some methods worked well only when the images are pre-registered and won't be practical if the pages have different alignments and different font-sizes, so we used image processing techniques and image matching algorithms to perform comparison of images . We followed an approach similar to to compare word images amongst two scanned pages

We find the unique corner points on the individual words and find a feature vector for these points. Then through clustering we assign an id to each point. Then we are able to use sequence of ids for a word to compare with other words for similarity. Then we take into account the context of the words to aid in finding the exact match for the words.

Integration into CritSpace

Peter Robinson notes that the greatest effect of the digital revolution is that it is empowering a new model of collaboration among scholars, and between scholars and readers. In sync with this, the goal of this project is to integrate the collation tool into CritSpace to greatly increase its usefulness. CritSpace [Figure 1] is a web based creativity support environment that implements spatial information management strategies to assist scholars in their open-ended research tasks.

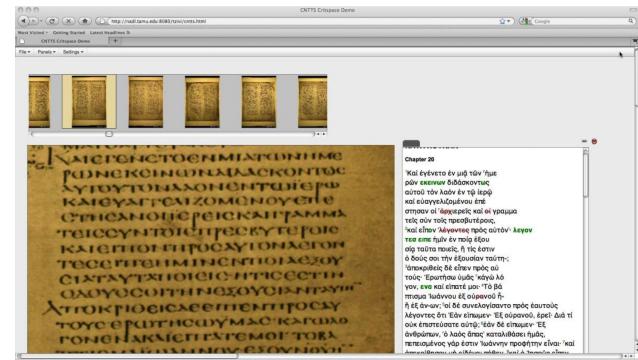


Fig. 1: Sample workspace with a text panel, image panel and facsimile viewer

The user-interface explained in figures below was planned to generate an effortless user-experience for digital scholars.

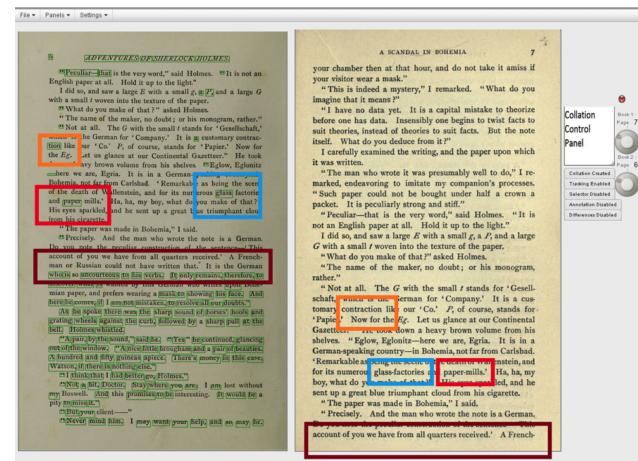


Fig. 2: Figure 2 Screenshot highlighting the differences with green boxes around the words. These are displayed when the user clicks on "Differences" button. Notable differences like missing hyphens are outlined as well as the end of the page.

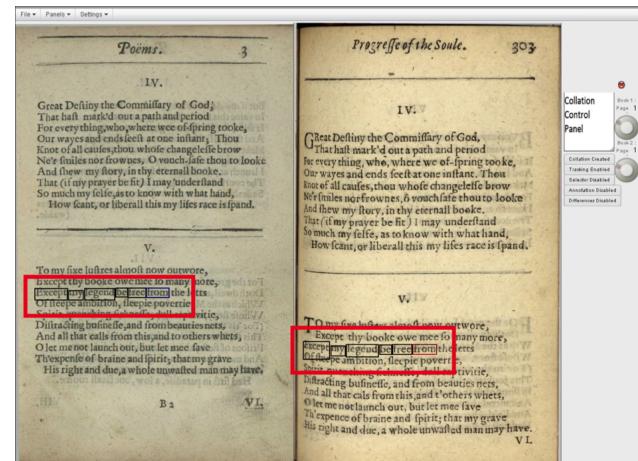


Fig. 3: Figure 3 Screenshot demonstrating the tracking feature. When the user hovers over any block of word its corresponding match is highlighted in the other page in red. The ones that have already been checked are turned black

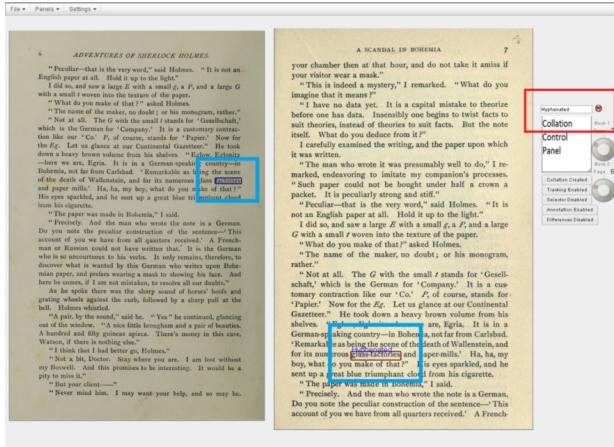


Fig. 4: Figure 4 Screenshot of the annotation feature. On enabling annotation mode, the user can select a word and a text box will appear. The text is displayed above the word every time annotation mode is set. A sample use-case has been outlined.

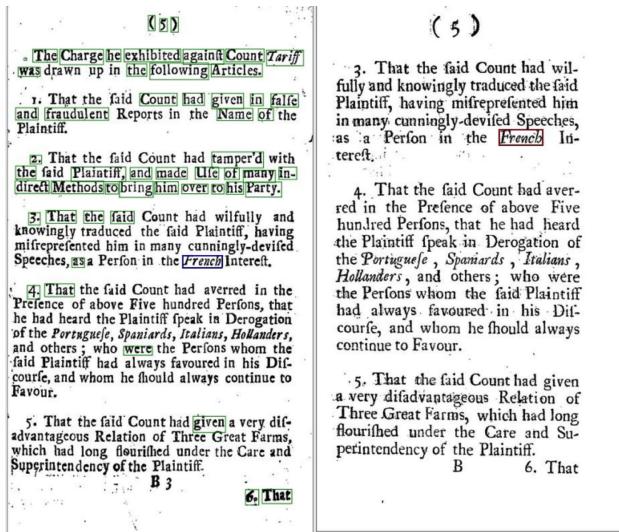


Fig. 5: Screenshot of collation output of two 17th century versions of The Late Tryal and conviction of Count Tariff



Fig. 6: Font variations in two versions of word "French"

Dataset

We tested the vHinman tool on various scanned texts available on the Internet Archive website and within TAMU collections. Some of them are digital copies of *Sherlock Holmes*, copies of early printings of Donne's Poems (1633 and 1635) and copies of The Late Tryal and conviction of Count Tariff. These books have many print and edition variants; for the pages of *Sherlock Holmes* tested, the tool shows an average accuracy of 95% in tracking the matches.

User Evaluation

Five subjects were chosen to participate in this study which was a mix of semi-structured interview regarding the experience of scholars on collation, followed by a demo of the prototype

and then by questions about the feedback of the tool and suggestions for its improvement.

ID	Area of interest	Career Stage
S1	Eighteenth century literature	Senior
S2	Bibliography	Senior
S3	Scholarly editing	Senior
S4	Scholarly editing	Senior
S5	Book history, Linguistics	Senior

Most of the subjects had prior experience with collation either in their scholarly research or for some classroom activities. Some of the subjects had experience with mechanical collators or text based collators. Many of the subjects still prefer the paper-based manual collation method because they find the supporting tools either inaccurate or too cumbersome to use or both. The need of collation in the subjects' research varied from the traditional scholarly editing process to bibliographic research and book history research.

S4 pointed out that he didn't have the resources to do the transcription for each of the documents he works on and also said that they are prone to errors.

S1 pointed out the need to be able to find differences in font-styles, ligatures like the move from using the long "s" to the current "s".

S2 liked the idea of integrating the vHinman into CritSpace which can foster collaborative work. She also liked the idea that the tool could have multiple panels (more than two). She pointed out that while supporting multiple images we can display the n-images in the form of medium sized thumbnails as is seen in "Google images", where the scholar can select any two panels to collate at a time. She noted that the tool could bring forward new uses of collation and could get collation adopted by scholars who currently don't focus much on it attributing the manual effort and inherent inaccuracies in the current method.

S5 suggested a novel use of the tool in verifying the authorship of a poem.

Some of the subjects felt the need to be able to point small differences like punctuation because this is important for a critical edition. Although our tool currently only supports identifying word differences, punctuation support can be added. S4 felt that the current implementation can quicken the collation process by addressing textual differences while punctuation can be addressed separately. The subjects in general liked the ability to use the original facsimile of the document via the tool rather than a transcription or a somewhat inaccurate OCR version of it.

Conclusion

This work has investigated the way humanities scholars perform collation work and what role collation plays in their research output. Collation is known to be a laborious and monotonous task unaided by technology so far. To address this problem, a prototype has been developed to perform collation in an automated manner. Image matching techniques are employed in building this prototype, so that the scholars can use the original facsimiles of the documents. The tool was integrated into CritSpace and will benefit from the collaborative environment. A user evaluation was conducted with experienced scholars. In summary, the tool looks very promising to the scholars and also has a high accuracy rate for the books tested so far. Thus this kind of tool can save a massive amount of time for scholars and set up a paradigm of digital collation encouraging even more scholars in finding new uses of collation in their work.

It extends the Hinman's principles by allowing collating multiple editions of a book in addition to multiple copies of same edition having minor differences.

Since it is has application in creating a critical edition, bibliography and book history research, this tool has the capability of gaining widespread adoption.

Future Work

Beyond printed material, it will be interesting to evaluate the tool for handwritten documents and make it robust for such documents. Also it will be great to test the tool for non-English documents. We can try out different visualization formats and different ways the scholars can use the output in their work. A detailed usability study can be conducted where scholars can perform some real collation work on few pages and compare their traditional method and the vHinman. Also the accuracy could be tested for warped images as most of the unobtrusive scanning methods produce some warping on the images.

References

- Unsworth J.** (2000) "Scholarly Primitives: what methods do humanities researchers have in common, and how might our tools reflect this?" In *Humanities Computing: formal methods, experimental practice* (13 May 2000)
- Schmidt, D., & Colomb, R.** (2009). "A data structure for representing multi-version texts online." *Journal of Human-Computer Studies*, 67(6), 497–514.
- Smith, S.E.** (2002) "Armadillos of Invention": A Census of Mechanical Collators, *Studies in Bibliography* 55, pp. 133-170
- Cream R.** "Sapheos Project", CDH University Of Southern Carolina sapheos.org
- Raabé W.** "Collation in Scholarly Editing: An Introduction" wraabe.wordpress.com/2008/07/26/collation-in-scholarly-editing-an-introduction-draft
www.juxtasoftware.org
collatex.net
- Lowe, D.G.** (2004) "Distinctive image features from scale-invariant keypoints." *Int. J. Comput. Vision*, 60:91–110, November 2004.
- Audenaert, N. and Furuta, R.** (2010) *What Humanists Want: How Scholars use Primary Source Documents*, Proceedings of the 10th Annual Joint Conference on Digital Libraries, pp. 283-292.
- Yalniz, I. and Manmatha, R.**, "An efficient framework for searching text in noisy document images," Proceedings of the 10th IAPR International Workshop on Document Analysis Systems (DAS'12)
- Robinson, Peter M. W.** (1994). "Collation, textual criticism, publication, and the computer." *Text* 7: 77-94. www.jstor.org/stable/pdfplus/30227694.pdf. 16 Dec. 2011
- Audenaert, N., Lucchese, G. and Furuta, R.** "CritSpace: a workspace for critical engagement within cultural heritage digital libraries" ECDL'10 Proceedings of the 14th European conference on Research and advanced technology for digital libraries
- Audenaert, N.** "CritSpace: An interactive visual interface to digital collections of cultural heritage material"
- Marshall, C.C., Shipman, F.M.** "Spatial hypertext and the practice of information triage." In: ACM Conference on Hypertext (Hypertext 1997), pp. 124–133 (1997)
- Rosten E., and Drummond T.** (2006) "Machine Learning for high-speed corner detection." In European Conference on Computer Vision. Volume I, 430-443, May 2006.

Swiss Voice App: A smartphone application for crowdsourcing Swiss German dialect data

Kolly, Marie-José
University of Zurich, Switzerland

Leemann, Adrian

University of Zurich, Switzerland

Dellwo, Volker
University of Zurich, Switzerland

Goldman, Jean-Philippe
University of Geneva, Switzerland

Hove, Ingrid
University of Zurich, Switzerland

Almajai, Ibrahim
University of Geneva, Switzerland

1 Introduction

The spatial variability found in dialects is an essential indexical property that is highly salient to listeners in everyday language situations: at social events, for example, one often hears conversations of the type "I have trouble localizing your dialect – where do you come from?". Although listeners are typically unaware of the underlying linguistic mechanisms involved, they are actively engaging in perceptual dialectology (cf. Preston 1989, Clopper & Pisoni 2004) and they seem keenly aware of dialectal variation. It is interesting then that different language speaking groups seem to recognize dialects of their language with different degrees of accuracy. Leemann & Siebenhaar (2008) and Guntern (2011) show that naïve Swiss German listeners can accurately recognize a speaker's dialect with a recognition rate of 86% and 74% respectively. However, Clopper & Pisoni (2005) report identification rates of only 30–50% for American and British English dialects; Kehrein, Lameli & Purschke (2011) report similar recognition rates for German dialects. Recent studies show that dialect recognition is possible via the mobile application *Dialäkt Äpp* (Leemann & Kolly, 2013; Kolly & Leemann, in review).

This contribution describes work in progress: *Voice Äpp*, currently in development at the University of Zurich, is a follow-up project on *Dialäkt Äpp*. The main purpose of both smartphone apps is to identify users' dialects on the basis of the dialectal variants of 16 words. *Dialäkt Äpp* users provide their pronunciation through tapping on the corresponding variant on the smartphone screen. However, the new *Voice Äpp* asks users to pronounce the word and uses automatic speech recognition (ASR) to identify users' pronunciation variants. The ASR training for *Voice Äpp* is partly based on acoustic data crowdsourced through *Dialäkt Äpp*. *Voice Äpp* further aims at illustrating the individuality in users' voices by providing a multidimensional profile of their voice. The launch of *Voice Äpp* is planned in December 2014.

Several research teams are interested in creating similar applications for other languages, using the frameworks put forth by *Dialäkt Äpp* and *Voice Äpp*: Mobile applications that recognize regional varieties of the entire German-speaking area, of American English, of British English, and of Italian, are currently under development.

2 Crowdsourcing data with Dialäkt Äpp

In 2013 we launched the iOS application *Dialäkt Äpp*, which capitalizes on the Swiss public interest in dialectology (Leemann & Kolly, 2013). We provided a functionality that, on the one hand, allows users to localize their own Swiss German dialect by indicating their pronunciation of 16 words (see Figure 1). Given the task to predict Swiss German dialects, a model was built by phoneticians who devised a set of maximally predictive words (i.e. maps from the Linguistic Atlas of German-speaking Switzerland: *Sprachatlas der Deutschen Schweiz* (SDS, 1962–2003)) that capture dialectal differences between localities. On the other hand, users can record their own dialect and listen to recordings of other users, thus discover the Swiss dialectal landscape. Figure 1 shows three screens of the application: the choice of dialectal variants for the word 'Donnerstag'/'Thursday', the identified localities as a list and on a map (Bern being the best hit in this example) and the

distribution of users' recordings covering German-speaking Switzerland.



Fig. 1: Screens of *Dialäkt Äpp*: (1) choice of dialectal variants with buttons; (2) result provided as a choice of five best hits and their corresponding positions on a map; (3) users' recordings (one pin per locality)

Dialäkt Äpp was launched on March 22, 2013, and has been downloaded over 58'000 times (as of February 28, 2013). The data recorded by this application contains (a) (written) choices of pronunciation for 16 words by each user who localized his/her dialect and (b) audio data for the same 16 words by each user who chose to record his/her voice. For (a), the corpus contains data from over 42'000 subjects (58% males, 42% females). Most users are from the cantons (and capitals) of Zurich, Bern, Basel, Luzern, Aargau, and St. Gallen. 64% of the users' pronunciation variants still correspond to the local variant recorded by the SDS (1962–2003) in the 1940's and a large number of users report that the localization of their dialect by the application is very close to their dialectal origin. For (b), the corpus counts 38'477 recorded variants stemming from a total number of 2'633 iOS devices (which corresponds roughly to the number of speakers; 54% males, 46% females). The geographical distribution of users corresponds to that of the data presented in (a).

The data elicited by *Dialäkt Äpp* has great potential for dialectological as well as forensic phonetic research. It can be used to create new dialect maps and compare them to the maps published in the SDS (1962–2003), thus to track sound change in progress. A number of maps have already been created (for the words *Apfelüberrest* 'apple core', *Bett* 'bed', *schneien* 'to snow', *Tanne* 'pine tree', and *tief* 'low'). Preliminary analyses show that phonetic isoglosses, as illustrated in maps like *Bett* (quality of /e/) and *Tanne* (quantity of /n/) are congruent with data from the SDS (1962–2003) (Kolly & Leemann, in review). The data can also be used to compare dialects at the acoustic phonetic level: For example, preliminary results show differences in speaking rate between the Bern dialect and the Zurich dialect (Leemann, Kolly, & Dellwo, accepted). Furthermore, this corpus can be used to create population statistics for a variety of phonetic parameters, which is desirable for forensic phonetic voice comparison (cf. Nolan et al., 2009).

3 Development of Voice Äpp

Voice Äpp has two major aims:

- To use ASR techniques to localize users' dialects
- To provide users with a multidimensional profile of their voice

3.1 ASR-based dialect localization

The novelty of this new project is to use ASR techniques instead of multiple choice buttons. Some difficulties can be expected as the ASR approach is not error-free, especially through a mobile application: recording conditions may vary a lot due to the distance from the microphone, noisy environments etc. However, the high-resolution microphones of smartphones, iPhones in particular, should facilitate the ASR task. Furthermore, identifying dialects, where small variation has to be taken into account, is not the initial purpose of ASR systems; the speech recognition domain aims at

normalizing such variation and at being rather dialect- or speaker-independent. In addition to this, the number of possible pronunciation variants for each word is important. For example, the word *Bett* 'bed' only counts two variants in the SDS (/bet/ and /bɛt/) whereas *Augen* 'eyes' has eleven dialectal pronunciation variants. The latter is highly discriminant – but the ASR task is more difficult. The algorithm will have to be modified since the voice recognition approach is not as reliable as the selection with buttons.

In order to achieve this, an ASR system is trained with two corpora: (a) the *Dialäkt Äpp* corpus described in 2 and (b) the TEVOID corpus (Dellwo, Leemann, & Kolly, 2012). Corpus (a) contains about six hours of speech of over 2'600 speakers, covering a dense net of local dialects in German-speaking Switzerland. Each recording is an isolated word from a set of 16 words. Corpus (b) contains two hours and 45 minutes of speech of 16 Zurich German speakers. Each recording is either a spontaneous or a read sentence. While the second corpus has been segmented by hand, the first one needs data preparation and verification as it was collected without control of linguistic content nor acoustic environment.

So far, encouraging results are obtained with limited training data. After ASR training with five variables from the *Dialäkt Äpp* corpus, dialect word recognition has reached accuracies of 92% (*Bett* 'bed'), 90% (*Kind* 'child', *Apfelüberrest* 'apple core'), 85% (*Tanne* 'fir tree'), 79% (*fragen* 'to ask'). These accuracies may increase with larger amounts of training data, which is currently being worked on.

3.2 Multidimensional voice profile and infotainment content

The second function of the *Voice Äpp* is a voice profile provided to the user. Based on a sentence recorded in their dialect, users learn about characteristics of their own voice in a playful way. A number of menus allow users to explore different aspects of speech, e.g. pitch, speech rate, articulation, auditory and visual perception.

Pitch: The fundamental frequency (f0) of the users' sentence is calculated and displayed in a histogram representing the distribution of the f0 of all the previous users.

Speech rate: The speech rate of the users' sentence is calculated and displayed in comparison to the previous users' speech rate.

Articulation: Users learn about sounds and their articulation. Upon clicking on an IPA symbol a sagittal cut is shown and the sound is played. In an interactive sagittal cut users move the position of the articulators and hear the corresponding vowel sound.

Auditory perception: Users can listen to what their sentence would sound like to a person with a hearing impairment/a cochlear implant.

Visual perception: Users are shown a video illustrating the *McGurk effect* (MacDonald & MacGurk, 1978) and the *Cocktail Party Effect* (Handel, 1989). Both effects illustrate that visual cues can be crucial for speech perception.

4 Conclusion

Voice Äpp should be as interactive as possible, allowing users to learn about the individual features of their dialect and their voice in a playful way. As shown by *Dialäkt Äpp*, a mobile application such as *Voice Äpp* is interesting for the user as well as for the researcher: by providing appealing content to the user, we gain large amounts of data. This crowdsourced data can be used to create population statistics, for example for analyses of speech prosodic features. In particular, *Voice Äpp* creates real time f0 and speaking time statistics, which represents a novelty for e.g. the field of forensic phonetics.

Acknowledgements

The project Swiss VoiceApp – Your voice. Your identity is funded by the Swiss National Science Foundation (SNSF); funding scheme: Agora; grant number: 145654.

References

- Clopper, C.G., & D. Pisoni** (2005). *Perception of dialect variation*. In: Pisoni, D., R.E. Remez (Eds.), *The Handbook of Speech Perception*, Oxford: Blackwell, 313–337.
- Dellwo V., Leemann, A., & Kolly, M.-J.** (2012). *Speaker idiosyncratic rhythmic features in the speech signal*. Proceedings of Interspeech2012. 9.-13.9.2012, Portland (OR), USA.
- Ferragne, E., & Pellegrino, F.** (2007). *Automatic dialect identification: A study of British English*. In: Speaker classification II. Berlin/Heidelberg, Springer: 243–257.
- Guntern, M.** (2011). *Erkennen von Dialekten anhand von gesprochenem Schweizerhochdeutsch*. Zeitschrift für Dialektologie und Linguistik 78/2: 155–187.
- Handel, S.** (1989). *Listening. An Introduction to the perception of auditory events*. MIT Press.
- Kehrein, R., Lameli, A., & Purschke, C.** (2010). *Stimuluseffekte und Sprachraumkonzepte*. In: Anders, C., Hundt, M., Lasch A. (Eds.). "Perceptual dialectology". Neue Wege der Dialektologie. Berlin/New York, de Gruyter: 351–384.
- Leemann, A., & Kolly, M.-J.** (2013). *Dialäkt Äpp*. <https://itunes.apple.com/ch/app/dialakt-app/id606559705?mt=8>.
- Kolly, M.-J. & Leemann, A.** (in review). *Dialäkt Äpp: Communicating dialectology to the public – crowdsourcing dialects from the public*. To appear in: Leemann, A., Kolly, M.-J., Schmid, S., & Dellwo, V. (Eds.). Trends in Phonetics in German-speaking Europe, Bern/Frankfurt: Peter Lang.
- Leemann, A., Kolly, M.-J., & Dellwo, V.** (accepted). *Crowdsourcing regional variation in speaking rate through the iOS app 'Dialäkt Äpp'*. To appear in: Proceedings of Speech Prosody 2014, 20.–23.05.2014, Dublin.
- Leemann, A., & Siebenhaar, B.** (2008). *Perception of Dialectal Prosody*. Proceedings of Interspeech 2008.
- MacDonald, John, & MacGurk, Harry** (1978). *Visual influence on speech perception processes*. Perception & Psychophysics 24/3: 253–257.
- Nolan, F., McDougall, K., de Jong, G., & Hudson, T.** (2009). *The DyViS database: style-controlled recordings of 100 homogenous speakers for forensic phonetic research*. The International Journal of Speech, Language and the Law 16/1: 31–57.
- SDS Sprachatlas der deutschen Schweiz.** (1962-2003). Bern (I-VI), Basel: Francke (VII-VIII).

Beautiful lips and porcelain cheeks: extracting physical descriptions from recent Dutch fiction

Koolen, Corina
c.w.koolen@uva.nl
University of Amsterdam, Netherlands, The

Wubben, Sander
s.wubben@uvt.nl
Tilburg University, Netherlands, The

van Cranenburgh, Andreas
andreas.van.cranenburgh@huygens.knaw.nl
University of Amsterdam, Netherland

1. Introduction

In literary analysis, description – as opposed to narration – has previously often been an underestimated part of fiction. Literary theorists such as Bal, Lopes and Nünning however have made a case for its relevance [1, 6, 8]. Lopes reviews how

well-known theorists like Barthes have dismissed description as 'extra', irrelevant or stalling the plot; he counters these notions with the statement that "[d]escription and narration constitute the two most basic modes of structuring any prose fiction text" [6, p. 19]. How the plot is conveyed, is relevant for how a text is judged. Literary theorist Wells for instance argues that description is the distinguishing factor between quality literature and 'simple' chick-lit novels [15]. Indeed, research has shown that literary novels contain significantly more noun phrases and prepositional phrases than chick lit, indicating a larger amount of description [5]. In this paper, the first steps are taken of a larger project in which description in fiction is computationally analyzed, as opposed to the now popular computational analysis of narrative (see for instance 7). The preliminary question that we want to answer is: how (well) can we extract descriptions from fiction? This will be tested in the current paper by zooming in on a specific domain: the physical description of fictional characters.

2. Motivation

Descriptions of physical appearance are chosen as a test case as they are more likely to occur in a current-day novel than for instance landscape description. Moreover, main characters are often introduced in the first chapters. This makes it possible in case of manual tagging (which we have done) to tag only the first chapters of a novel. Finally, it would be an interesting feature for further literary interpretation. Connotations of beauty in folk tales have been researched [i.e. 14], but this has not yet been done for novels.

3. Method

The corpus of [5] is used, consisting of 32 novels of recent Dutch fiction, half chick-lit, half literary novels. Two of them were tagged from beginning to end for descriptions of physicality, including clothing. One is a literary novel, *De schilder en het meisje* ('The painter and the girl') by Margriet de Moor, the other chick lit, *Zwaar verliefd* ('Heavily in Love') by Chantal van Gastel. Bal defines description as "a textual fragment in which features are attributed to objects" [1, p. 36], a definition we will follow. We tagged full sentences that were either mainly concerned with physical appearance (example 1a, Van Gastel), mentioned a single feature (1b, De Moor) or somewhere in between.

1a. Hij heeft mooie lippen. *He has beautiful lips*.

1b. Door de rook heen keek hij naar de porseleinen wangen van mevrouw Cloeck[.] *Through the smoke he watched madam Cloeck's porcelain cheeks[.]*

For the extraction, two approaches are compared: (1) manual development of lexical-linguistic patterns and (2) a Naive Bayes and an SVM classifier. For the former, because patterns were manually developed on the basis of two novels, the patterns were subsequently tested on the other 30 novels, each of which the first 500 sentences were manually tagged.

3.1 Lexical-linguistic patterns

After an initial exploration of the two main novels' tagged sentences, an approach was adopted of manually developing patterns to detect sentences containing description. Hearst uses similar patterns to harvest hyponyms [3]. Patterns consist of a combination of linguistic and lexical information, see example 2 below. A set of 13 patterns was written. The manual exploration showed that sentences containing physical descriptions, as opposed to sentences with no such descriptions, (a) contain more nouns and adjectives, (b) are regularly coupled with a few specific, static verbs, and (c) contain a couple of recurring base lexical-linguistic patterns, e.g., 'He was [a manNP] [[withPP] [brown eyesNP]]'. To perform extraction, the corpus was parsed with Dutch parser Alpino [2, 12]. Alpino parse trees provide rich linguistic annotations of sentences such as grammatical function of constituents. The trees can be queried with XPath, which was integrated in Van

Cranenburgh's TreeSearch interface [5]. Linguistic information alone does not suffice however to target physical descriptions, so we used Cornetto, the Dutch WordNet [13], to expand a manually constructed lexicon of nouns and adjectives related to physical descriptions. The lists were cleaned to exclude words that were not relevant to the topic, resulting in a lexicon of almost 600 words.

An example of a pattern translated to an xPath query is:

```
//node[@cat="pp" and @rel="mod"]//node[%uiterlijkA%]../
node[%uiterlijkN% or %kleding%]
```

Example 2: This pattern searches for a modifying prepositional phrase which contains an adjective and a noun from the lexicon.

3.2 Machine learning

We cast the task of extracting physical descriptions as a text classification task in order to use machine learning methods. The task then becomes for a given text to automatically assign a class to it (in our case: physical description or no physical description). Usually, text classification is done on the document level. This means that for each document a corresponding class is predicted [10]. Algorithmic methods used for the classification task vary widely. Naive Bayes classification and Support Vector Machines (SVM) were used, two established straight-forward approaches to text classification [4, 9, 11]. We adapted these approaches to our task of classifying sentences. Each sentence was classified as either a description or not, in order to extract the descriptions.

4. Results

4.1. Lexical-linguistic patterns

Precision, recall and F-measure were calculated for each pattern separately for the two main novels, for the test set of 30 novels, for a cumulative set of all pattern results, and for chick-lit versus literary novels; the most important results can be found in table 1. Sentences that were extracted more than once were calculated as one hit.

4.1. Lexical-linguistic patterns

	F-measure (%)	Precision (%)	Recall (%)
Test set-all novels	31	29	35
Test set-literature	25	29	22
Test set-chick lit	18	28	13
Main novels	16	24	12

Table 1: Results for lexical-linguistic pattern-based extraction

An unexpected outcome was that the results were much better for the 30 novels in the test set than for the two novels on the basis of which the patterns were developed; the percentage of descriptions might be higher in the first chapters. Another interesting result was the performance on literary novels, which was better than on chick lit. An explanation might be that in chick lit, sentences are shorter [see 5], more often elliptic ('And his mouth... He has beautiful lips. Precisely full enough.') and regularly discuss physicality through dialogue, for which it is hard to develop patterns. Generic patterns, containing little more than lexical information, achieved higher scores than more specific ones. The specific patterns did improve

the cumulative outcome. Further research is needed, but an expansion of the lexicon might raise performance.

4.2 Machine learning

We trained our classifiers on the two annotated novels. The features selected as input for the classifiers are words weighted with tf.idf, for which we considered the sentences as documents and the novels as the collection of documents. Experiments were also performed for bigrams and part-of-speech tags, but the results were comparable to the results we report here. We performed ten-fold cross validation on the set of sentences from each novel and both novels combined. We found that Naive Bayes outperforms SVM for this task, as can be observed in Table 2.

	F-measure (%)	Precision (%)	Recall (%)
Both novels			
Naive Bayes	60	57	62
SVM	58	59	58
Zwaar verliefd			
Naive Bayes	62	61	64
SVM	57	59	56
De schilder en het meisje			
Naive Bayes	58	55	62
SVM	52	53	51

Table 2: Results for the Naive Bayes and SVM classifier

Performance is considerably higher than that of the pattern-based approach. The skewedness of the class distribution (descriptions form only a small portion of a novel) makes this classification task a hard one, but overall this is a promising method. This machine learning approach can be regarded as a baseline: more sophisticated methods might yield better results.

5. Conclusion

A comparison of two methods for extracting sentences containing physical descriptions paints a clear picture: extracting such information is a complex matter but not impossible, and machine learning performs better than a manual-based approach. However, the main benefit of using the manual tagging and patterns is the insight they give in the form of the sentences that contain the sought-after descriptions, whereas the bag-of-words approach of the machine learning method is limited to finding features based on individual words. A possibility for future research is extension of the patterns and the lexicon to see if the results can be improved, but we prefer to pursue a bottom-up approach. A combination of the methods could be fruitful: using the patterns as features for machine learning. We could also explore descriptions on a different textual level; especially for the chick-lit novels, where use of ellipsis and dialogue confuses sentence extraction, larger fragments of texts should be analyzed. Targeted topic modeling might be useful for this purpose.

References

1. **Bal, M.** (2009). *Narratology: Introduction to the Theory of Narrative*. Toronto: University of Toronto Press.
2. **Bouma, G., Van Noord, G. and Malouf, R.** (2001). *Alpino: Wide-coverage computational analysis of Dutch*. Language and Computers, 37(1). 45–59.
3. **Hearst, M.A.** (1992). *Automatic Acquisition of Hyponyms from Large Text Corpora*. In Proceedings of the 14th Conference on Computational Linguistics, vol. 2 .539–545.

4. Hearst, M.A., Dumais, S. T., Osman, E., Platt, J., and Scholkopf, B. (1998). *Support vector machines*. Intelligent Systems and their Applications, IEEE, 13(4). 18-28.
5. Jautze, K., Koolen, C., Van Cranenburgh, A. and De Jong, H. (2013). *From High Heels to Weed Attics: a Syntactic Investigation of Chick Lit and Literature*. In Proceedings of the Second Workshop on Computational Linguistics for Literature. <http://aclweb.org/anthology//WW13/W13-1410.pdf>.
6. Lopes, J. M. (1995). *Foregrounded Description in Prose Fiction: Five Cross-literary Studies*. Toronto: University of Toronto Press.
7. Mani, I. (2013). *Computational Narratology*. In The Living Handbook of Narratology. Eds. Hünn, P., Schmid W. and Schönert, J. <http://www.lhn.uni-hamburg.de>.
8. Nünning, A. (2007). *Towards a Typology, Poetics and History of Description in Fiction*. In Description in Literature and Other Media. Eds. Wolf W. and Bernhart W. Amsterdam, New York: Rodopi. 91–128.
9. Rish, I. (2001). *An empirical study of the naive Bayes classifier*. In IJCAI 2001 workshop on empirical methods in artificial intelligence, 3(22). 41-46.
10. Sebastiani, F. (2002). *Machine learning in automated text categorization*. In ACM computing surveys (CSUR), 34(1). 1-47.
11. Steinwart, I. and Christmann, A. (2008). *Support vector machines*. New York: Springer.
12. Van Noord, G. (2006). *At last parsing is now operational*. In TALN06. Verbum Ex Machina. Actes de la 13e conference sur le traitement automatique des langues naturelles. 20–42.
13. Vossen, P., Hofmann, K., De Rijke, M., Tjong Kim Sang, E., and Deschacht, K. (2007). *The Cornetto Database: Architecture and User-scenarios*. In Proceedings of the Dutch-Belgian Information Retrieval Workshop. 89–96.
14. Weingart, S. and Jorgensen, J. (2012). *Computational Analysis of the Body in European Fairy Tales*. Literary and Linguistic Computing, 28(4).
15. Wells, Juliette. (2005). *Mothers of Chick Lit? Women Writers, Readers, and Literary History*. In Chick Lit: The New Woman's Fiction. Eds. Ferriss S. and Young, M. New York: Routledge. 45–70.

TheoPhilo. A prototype for a Thesaurus of Philosophy

Lamarra, Antonio

antonio.lamarra@cnr.it

CNR- Istituto per il Lessico Intellettuale Europeo e Storia delle Idee

Tardella, Michela

michela.tardella@iliesi.cnr.it

CNR- Istituto per il Lessico Intellettuale Europeo e Storia delle Idee

1. Introduction

Our paper aims at presenting TheoPhilo-Thesaurus of Philosophy, the prototype of a digital multilingual thesaurus in the field of Philosophy. With 'thesaurus' we mean a concept-based collection of terms¹ that we are building by means of philosophical texts and dictionaries.

The purpose of TheoPhilo is to test potentials and limits opened by the interaction between digital tools and devices (such as digital archives and libraries) and the long tradition of historical and lexicological studies of the *Istituto per il Lessico Intellettuale Europeo e Storia delle Idee* (www.iliesi.cnr.it), which developed the thesaurus. The work on the prototype implied the cooperation of experts in History of philosophy, Linguistics and Computer science, and has been conceived in the frame of the semantic web. We are currently testing the collection of terms and building up an ontology finalized to semantic enrichment and to information retrieval of the digital resources uploaded in the portal *Daphnet. Digital Archives of Philosophical Texts on the NET* (www.daphnet.org).

The plurality of languages coexisting in the portal lead us to design a multilingual collection of terms, in order to enable scholars, students, teachers and other interested users to search within the large quantity of texts in *Daphnet* and – this being an important added value – to make queries by using one's own mother tongue. TheoPhilo is not a dictionary of philosophy and it does not offer definitions of philosophical terms: we want it to be a useful map to enrich texts and to retrieve information.

TheoPhilo itself could be an interesting object of research as, once completed, it will allow a linguistic analysis of the philosophical terminology structured according to the criteria we are adopting (see § 3). For this reason we intend to carry out a parallel work on this tool, developing a linguistic and an historical-philosophical study on the collection of terms we have selected.

2. The content

The *Daphnet* portal, implemented within the project AGORA. Scholarly Open Access Research in European Philosophy (www.project-agora.org)^{2 3}, consists of two Open Access platforms, *Ancient and Modern Philosophy*. The first one contains (a) the transcription of the collection of Presocratic thinkers originally edited by H. Diels and W. Kranz, with the Italian translation edited by G. Giannantoni; (b) the transcription of the *Socratis et Socratiorum Reliquiae* by G. Giannantoni; (c) the volume *Vita e opinioni dei filosofi* (the editorial collection is by R. D. Hicks, H. S. Long and M. Marcovich, the Italian translation is by M. Gigante); (d) the *Opera Omnia* of Sextus Empiricus (ed. by H. Mutschmann). The second platform, *Modern Philosophy*, gives access to a number of Latin, Italian, French and German texts which are considered representative of the philosophical thinking of the 16th, 17th and 18th centuries. It includes works by Alexander Gottlieb Baumgarten, Giordano Bruno, René Descartes, Immanuel Kant, Gottfried Wilhelm Leibniz, John Locke, Baruch Spinoza and Giambattista Vico. *Daphnet* also presents an OJS platform dedicated to secondary sources, the *Daphnet Digital Library*, containing a wide selection of articles published in the journal *Elenchos. Rivista di studi sul pensiero antico* (Bibliopolis, Napoli), in *Lexicon Philosophicum. Quaderni di Terminologia filosofica e storia delle idee*, and in the volumes dedicated to the proceedings of the international conferences organized by CNR-ILIESI. In addition to these critical essays, the portal presents the monograph by Emidio Spinelli, *Questioni Scettiche: letture introduttive al pirronismo antico* (Lithos, 2005) and the brand new online journal *Lexicon Philosophicum: International Journal for the History of Texts and Ideas* (www.lexicon.cnr.it).

The set of lexical items extracted from texts and multilingual philosophical dictionaries (see § 3) includes Nouns, Adjectives, Verbs and Adverbs, both monorhematic and multiword expressions. Fig. 1 shows a query's result, namely for the French philosophical subject 'abduction'. The software presents the interlinguistic equivalents available in the five languages implemented so far (Latin, Greek, Italian, French, and English). Fig. 2 shows another example, related to the multiword expression 'harmonie préitable', that is presented in TheoPhilo both as autonomous lexeme included in the alphabetical list and in the box related to its belonging terms (Fig. 3). As a result, from the expression 'harmonie préitable', it is possible to reach the box related to the belonging lexeme 'harmonie' and viceversa.

TheofPhilo Thesaurus of Philosophy

Cerca: Lingua: Invia RESET

Entrata: abduction
Lingua: FRANCESE
Note: Acquisizione ex novo da MASO

Traduzioni

GRECO:	ἀπαρτίνι
INGLESE:	abduction
ITALIANO:	abduzione
LATINO:	abductio reductio

Modifica Elimina

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

Lessico Intellettuale Europeo e Storia delle Idee - CNR © 2013 ILIESI. Tutti i diritti riservati
Progetto Agora © 2010-2013

Fig. 1: Philosophical Subject 'abduction'

TheofPhilo Thesaurus of Philosophy

Cerca: Lingua: Invia RESET

Entrata: harmonie
Lingua: FRANCESE
Note: Greco tratto da ABB.

Sintagmi
harmonie préétablie

Traduzioni

GRECO:	ἀρμονία
INGLESE:	harmony
ITALIANO:	armonia
LATINO:	concentus concordia harmonia

Modifica Elimina

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

Lessico Intellettuale Europeo e Storia delle Idee - CNR © 2013 ILIESI. Tutti i diritti riservati
Progetto Agora © 2010-2013

Fig. 3: Philosophical Subject 'harmonie'

TheofPhilo Thesaurus of Philosophy

Cerca: Lingua: Invia RESET

Entrata: harmonie préétablie
Lingua: FRANCESE
Note: Sinonimo: accord préétabli

Composta da
harmonie

Traduzioni

INGLESE:	preestablished harmony
ITALIANO:	armonia prestabilita
LATINO:	harmonia praestabilita

Modifica Elimina

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

Lessico Intellettuale Europeo e Storia delle Idee - CNR © 2013 ILIESI. Tutti i diritti riservati
Progetto Agora © 2010-2013

Fig. 2: Philosophical Subject 'harmonie préétablie'

TheofPhilo Thesaurus of Philosophy

Cerca: Lingua: Invia RESET

Entrata: acte
Lingua: FRANCESE

Traduzioni

GRECO:	ἐντελέχεια ἐνέργεια
INGLESE:	act action activity
ITALIANO:	atto
LATINO:	actio actus

Modifica Elimina

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

Lessico Intellettuale Europeo e Storia delle Idee - CNR © 2013 ILIESI. Tutti i diritti riservati
Progetto Agora © 2010-2013

Fig. 4: Philosophical Subject 'acte'

3. Tools and procedures

The ontology^{4 5 6} built to represent the content, is structured according to the following four categories: *Persons* (philosophers, scholars); *Relevant Concepts* (philosophical

subjects, relevant events); *Relevant Subjects* (geographical entities, philosophical themes, philosophical schools, quotations, titles); *Sources* (secondary and primary sources).

TheofPhilo's specific purpose is to populate the sub-category of the philosophical subjects by two typologies of relations: interlinguistic equivalence and intralinguistic semantic relations.

From the procedural point of view, the work was carried out according to the following phases: digitization in Excel format of the multilingual entries systems lemmatized in N. Abbagnano's *Dizionario di Filosofia* (Torino 1998) and in A. Lalande's *Vocabulaire technique et critique de la philosophie* (Paris 1983); merging of the Greek and French philosophical subjects selected for the semantic enrichment experiments; acquisition of relevant French and Greek terminology using S. Maso's *Lingua Philosophica Graeca* (Milano-Udine, 2010), in which relevant Greek philosophical terms are presented along with their Latin, Italian, English, French and German equivalents; acquisition of Latin, Italian and English equivalents. In the last phase, in order to enrich the interlinguistic equivalences, we used the following lexicographical sources: A. Bailly, *Dictionnaire Grec-Français*, Édition revue par L. Séchan et P. Chantraine, Paris 1950 (16th ed.); J. M. Baldwin, *Dictionary of Philosophy and Psychology*, Gloucester, Mass. 1960 (2nd ed.); B. Cassin, *Vocabulaire Européen des Philosophie*, Tours 2004; *Enciclopedia filosofica*, Roma 1979 (2nd ed.); L. Rocci, *Vocabolario Greco-Italiano*, Perugia 1993 (37th ed.); Liddell-Scott, *Greek-English Lexicon*, Rev. by H. S. Jones, Oxford 1968 (9th ed.); T. Sanesi, *Vocabolario Italiano-Greco*, Pistoia-Siena 1916 (12th ed.).

Currently the philosophical subjects are being implemented in a relational MySQL database created and managed by Dr. Ada Russo. This technology guarantees a more efficient data management, helps to control interlinguistic relation (equivalence), and supports the acquisition of Latin, Italian and English Subjects. At present the number of terms consists of 4549 (1006 Greeks, 948 French, 909 Italian, 895 English, 791 Latin), but it will increase, we in fact intend to implement also Spanish and German philosophical terminology.

In order to build the ontology, we have been using *Pundit* (<http://thepund.it>), a semantic web annotator created by Net7 (www.netseven.it) and employed in the semantic enrichment and semantic interlinking activities in the frame of the AGORA project. Conceived in the increasingly wider context of the semantic web technologies for encoding, managing and enriching digital object, *Pundit* allows to produce semantic annotations, whose semantics is machine-processable. Each annotation consists of a RDF triple, which is a statement made up of a subject (S), a predicate (P) and an object (O)⁸; according to this technology you could create, for example, a triple that states: "Sextus Empiricus (S) is the author of (P) *Pyrrhoneae Hypotyposes* (O)".

Among the variety of annotations, the most useful to our purposes are those implying triples which connect:

- A textual fragment to a philosophical subject, such as:
“The selected text fragment (S) *DealsWith* (P) the Greek philosophical subject *to alethes* (O)” (Fig. 5) or “The selected text fragment (S) *Defines* (P) the Greek philosophical subject *ataraxia* (O)” (Fig. 6). According to its project’s goals, CNR-ILIESI team preconfigured the following properties (the inverse properties in brackets):
 1. *Defines* (*IsDefinedBy*);
 2. *IndirectlyDefines* (*IsIndirectlyDefinedBy*);
 3. *IsAnExtensionalInstanceOf* (*IsExtensionallyInstantiatedBy*);
 4. *IsAnIntensionalInstanceOf* (*IsIntensionallyInstantiatedBy*);
 5. *DealsWith* (*IsDealtWithBy*).
 - A philosophical subject to another philosophical subject. While the interlinguistic equivalents are already available in TheoPhilo (see §2 and related figures), the intralinguistic relations will be implemented during the next phase of our work. RDF triples will be created considering philosophical subjects both as Subject and as Object of each triple and connected by the following properties (the inverse properties in brackets):
 1. *IsSynonymOf* (*HasSynonym*);
 2. *IsHomonymOf* (*HasHomonym*);
 3. *IsHyperonymOf* (*HasHyperonvm*);

4. *IsHyponymOf* (*HasHyponym*);
 5. *IsCo-HyponymOf* (*HasCo-Hyponym*);
 6. *IsAntonymOf* (*HasAntonym*).

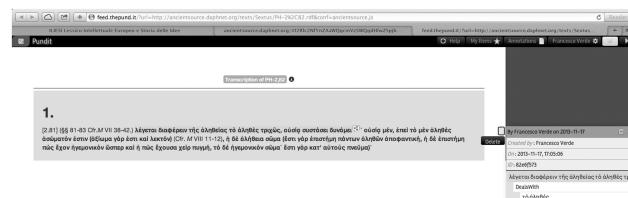


Fig. 5: Annotation Text to Subject with the RDF property `DealsWith`

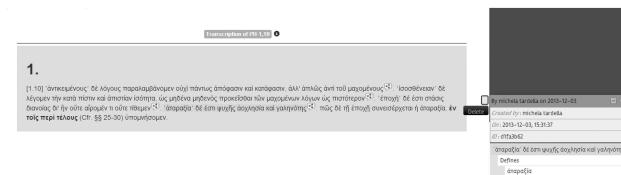


Fig. 6: Annotation Text to Subject with the RDF property `Defines`

4. Conclusions

TheoPhilo is a tool still in its pilot phase and work at its implementation is in progress. However, once completed the semantic annotation activity, it will enable users to access texts dealing with the subject the query was made for, according to the specific language chosen by the user. TheoPhilo will also allow scholars to deepen their research, making queries on the texts, according to both interlinguistic and intralinguistic relations. Furthermore, TheoPhilo will be considered as a large corpus of philosophical -interweave- terms (around 5000 at present) and it will be analysed from both linguistic and historical-philosophical approach.

References

1. **Magris, M. et al.** (2002). *Manuale di terminologia*. Hoepli, Milano.

2. **Tardella, M.** (2013). Agora. Scholarly Open Access Research in European Philosophy, Blityri. Studi di storia delle idee sui segni e le lingue, II (1): 167-174.

3. **Marras C. and Lamarra A.** (2013). *Scholarly Open Access Research in Philosophy: Limits and Horizons of a European Innovative Project*, Proceedings of Digital Humanities International Conference 2013 (University of Nebraska-Lincoln 7/13-15/2013). dh2013.unl.edu/abstracts/ab-316.html (date accessed: 10/31/2013).

4. **Garshol, L. M.** (2004), *Metadata? Thesauri? Taxonomies? Topic Maps! Making sense of it all*, 6. www.ontopia.net/topicmaps/materials/tm-vs-thesauri.html#N773 (date accessed: 10/31/2013).

5. **Grenon P. and Smith B.** (2009), *Foundations of an ontology of philosophy*, Synthese (2011): 185–204.5.

6. **Gruber, T. R.** (1993). *A translation approach to portable ontology specification*, Knowledge Acquisition, 5 (2):4. 199-220

7. **Andrews, P. et al.** (2011). *A classification of semantic annotation systems*, Semantic Web Journal, 0 (2011): 1-27.7. www.semantic-web-journal.net/sites/default/files/swj123_0.pdf (date accessed: 10/31/2013).

8. **Grassi, M. et al.** (2013). *Pundit: augmenting web contents with semantics*. Literary and Linguist Computing, 288. (4): 640-659 (date accessed: 02/20/2014).

The social pleasure of the text: Applying digital humanities methods to reception studies

Lang, Anouk

anouk.lang@gmail.com

University of Strathclyde

How do readers use social media to express the value and the pleasure that the experience of reading holds for them? And, given the rapidity with which corpora gathered from social media are growing, what kinds of methods are most useful for analysing this kind of (big) data so as to cast light on the phenomenology of reading experiences? This paper seeks to answer these questions by presenting the findings of a project on developing methods for analysing and evaluating literary engagement in digital contexts, funded by the Arts and Humanities Research Council under the auspices of the Cultural Value Project.¹ It will report on what can be learnt from the large amount of user-generated data available on microblogging services and social network sites about the value that reading brings to the lives of individuals and communities, and will offer an evaluation of the various analytical tools and methods available to scholars working on reading and reception studies who wish to include born-digital data in their research.

Work in reception studies is increasingly focusing on the ways that an understanding of the significance of individual reading experiences can be enriched by attending to occasions when readers join with others to express opinions about a text, and work together to construct its meaning. Scholars have argued that it is in fact in these acts of *public* negotiation of meaning – for example book group discussions – that readers can be observed doing the *private* cognitive work of textual engagement, as their interpretations change in the act of articulating their response in a social context.² The fact that the rich textual data available on social media is often generated by readers in conversation with friends or acquaintances, in contexts quite different to interviews with researchers or questionnaires which might prompt a higher level of self-editing, makes it even more compelling to work with.³ The obvious advantage of working with this sort of born-digital material is that it lends itself to analysis using the growing number of tools and methods being developed within digital humanities, which have the power to integrate textual and geospatial information, and to identify lexical trends in time-stamped data. Such computational methods not only offer scholars the opportunity to analyse much larger bodies of text than is ordinarily possible for individual researchers to examine through close reading, but also to draw on, and discover patterns in, temporal and geospatial metadata.

Data for this project was gathered from two different social media platforms, the microblogging platform Twitter and the book collection website LibraryThing.⁴ For the Twitter data, searches were performed for literary prizes (for example *Man Booker Prize* and *Nobel*), author names (for example [Eleanor] Catton and [Alice] Munro), and hashtags commonly used to signal reading-related tweets (for example #goodreads and #mustread). For the LibraryThing data, the results of the Twitter searches were used to suggest particular books to investigate, so as to enable a comparison of the way readers discussed books on the two platforms. The numerical review scores and the text of user reviews of these books were stored in a database, along with metadata about the user. While some interesting work on literary value has already been done by scraping data from Amazon,⁵ LibraryThing was selected for this project as it is a platform where readers gather primarily to share information voluntarily about books in ways not (directly) linked to commercial activity. Moreover, it is also possible to link some of this information to users' reported geographic location, something which cannot be done with Amazon data.

Various digital methods were then applied to the resulting datasets: thematic analysis using methods from corpus linguistics, analysis of trends in word usage over time using a burst detection algorithm, and geospatial analysis.

1) Thematic analysis

Analytical techniques from corpus linguistics were employed to identify patterns of unusually prominent words, phrases and grammatical constructions. The textual data gathered

were tagged with the CLAWS part-of-speech tagger,⁶ and the concordance program AntConc⁷ was then used to identify the most frequent words, determine their statistical significance as compared to a reference corpus, find the terms that most commonly collocated with them, and carry out other analytical procedures. Sub-corpora were separated out by hashtag and geographical location, and analysed individually.

2) Temporal analysis

As all the Twitter data and a significant proportion of the LibraryThing data is time-stamped, it presented an opportunity to analyse trends over time, something that can be done with burst detection analysis in order to gauge how influential particular words or hashtags have been over time.⁸ The Sci2 tool⁹ was used to perform burst detection, and to visualise the results as temporal bar graphs. Terms that "burst" into prominence were then fed back into the corpus linguistic analysis, for example in order to examine the collocation patterns around them, and to attend to the context in which they initially appeared.

3) Geospatial analysis

The software package ArcGIS was used to create a GIS database including layers derived from the Twitter and LibraryThing data, to see where particular geographical patternings in the search terms and hashtags occurred. (While not all tweets or contributions to LibraryThing have georeferences attached to them, a large enough number do to make this form of analysis worthwhile.) These data were then layered against census data (such as level of educational attainment or socioeconomic status) aggregated at the output area level, in order to enable semantic patterns in the articulation of reading-related tweets and posts to be considered alongside the demographic features of the places where they were articulated.

The paper will set out the advantages offered by thematic, temporal and geospatial analyses, and suggest the components of cultural value which are best addressed by each, while also considering how these different forms of analysis may be productively combined.

References

1. www.ahrc.ac.uk/Funded-Research/Funded-themes-and-programmes/Cultural-Value-Project/
2. Daniel Allington and Bethan Benwell (2012), *Reading the Reading Experience: An Ethnomethodological Approach to 'Booktalk'*, in From Codex to Hypertext: Reading at the Turn of the Twenty-first Century, ed. by Anouk Lang (Amherst, MA: University of Massachusetts Press, 2012), pp. 217–233.
3. Rhiannon Bury, Ruth Deller and Adam Greenwood (2013), *From Usenet to Tumblr: The Changing Role of Social Media*, *Participations* 10, 299–318.
4. <https://twitter.com/>; www.librarything.com/.
5. Ed Finn (2011), *Becoming Yourself: The Afterlife of Reception*, Pamphlets of the Stanford Literary Lab 3. 15 Sept 2011. litlab.stanford.edu/?page_id=255. 1 Nov 2013.
6. ucrel.lancs.ac.uk/claws/.
7. www.antlab.sci.waseda.ac.jp/software.html.
8. Jon Kleinberg b(2002), *Bursty and Hierarchical Structure in Streams*, Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '02 (New York: ACM, 2002), pp. 91–101.
9. sci2.cns.iu.edu.

BFM Collection - Open-Source Digital Editions of Medieval French Texts

Lavrentiev, Alexei

alexei.lavrentev@ens-lyon.fr
ICAR - CNRS

1. Introduction

The project of the BFM collection of digital edition was born as a result of over 20 years of developing a large corpus of medieval French texts for research purposes. This corpus called *Base de français médiéval* (BFM, <http://txm.bfm-corpus.org>) currently includes over 130 texts, totalling approximately 4.7 million of words. The essential part of the corpus is composed of digitized paper scholarly editions selected for their philological quality. Digitizing paper editions was the only way to build a substantial corpus in a relatively short time and with limited funding. However there are serious copyright issues related to printed editions, as publishers tend to require exclusive rights on the books they print, and they are often reluctant to authorize digitization and re-use of the data in text corpora. Before 2000, publishing contracts rarely included explicit clause on digital distribution, so it can be argued that scholarly editors (or their heirs) still hold copyright for this medium, but more recent contracts include long lists of digital products and distribution modes. Even though possibilities for open-licensed publishing on the web exist, scholars have to give up all their rights if they want to publish their works in a prestigious collection recognized by the academic community. The BFM team aims at providing scholars with a possibility to publish medieval French texts under an open license (like CC BY-SA) in a collection with editorial quality guaranteed by the expertise of the reading committee including leading specialists in medieval French language and literature, in text editing techniques and in digital philology.

2. Editing principles

In addition to the open licensing, the BFM collection marks itself out by innovative editing principles. These principles have been elaborated in the project of the *Queste del saint Graal* digital edition¹ and include a multi-layer transcription of primary sources (at least normalized and diplomatic), particular attention to punctuation and word segmentation, careful and clearly marked correction of scribal errors, linguistic annotation (part-of-speech and direct speech tagging). The "bedierist" method of the "best witness" is generally applied, but transcriptions of additional aligned witnesses are encouraged. Whenever possible, an edition should include a digital facsimile of the primary source which allows verifying the quality of the transcription. The presence of a modern French translation is optional but may be very useful to increase the range of potential readers and uses. All these principles are described in detail in the Introduction to the *Queste del saint Graal* edition and most of them were presented and discussed at the International Congress of Romance Linguistics and Philology (CILPR) in 2013².

3. Workflow and publication platform

At the first stages of the editing process text editors like Microsoft Word or Libre Office Writer may be used for the convenience of scholarly editors. A small number of special characters and character or paragraph styles are defined to facilitate future processing. For instance, a hash symbol before a letter indicates that a small letter from the primary source should be capitalized in the normalized transcription. Once the primary editing complete, the text is converted to XML-TEI, which is the pivot format for all markup and editorial

products in the BFM collection. The BFM corpus preparation chain automatic tools for tokenization, morphosyntactic annotation and direct speech markup. Whenever possible, the morphosyntactic annotation is verified by experts, as the automatic tagging of Old French produces inevitably a certain number of errors due to the high level of orthographic and morphological variation.

The BFM web portal built on the TXM platform³ will be used to publish the collection on the web. The advantage of this portal is that it combines the possibility to render the edition in a convenient form for reading (including parallel browsing of multiple transcription layers, digital facsimile and translation) with powerful tools for qualitative and quantitative text analysis (including frequency lists, KWIC concordances, specificity, factorial analysis, etc.). The editions of the BFM collection will be included in the BFM main corpus, and the BFM registered users will benefit from additional services, such as creating a subcorpus or recording queries. However, the possibility to read the text and to download XML-TEI source files or a PDF printable version will be provided without registration requirement.

4. Current state of the project

The edition of the *Queste del saint Graal* is currently complete from the philological point of view (although additional manuscript transcriptions will probably be produced in the future). All the major components of this edition (multiple layer transcriptions, modern French translation, manuscript images, introduction, proper name index and glossary) are available on the BFM portal (<http://txm.bfm-corpus.org>) through a special "GRAAL" corpus. However, a more convenient interface for browsing the edition and for direct access to its components is still under development.

More editions are being prepared at a more or less advanced stage. These include the *Psautier d'Arundel* (edited by C. Pignatelli and A. Lavrentiev)⁴, the *Vie de saint Alexis* (edited by C. Marchello-Nizia and T. Rainsford) and the first French texts, *Serments de Strasbourg* and *Séquence de sainte Eulalie* (edited by C. Guillot, A. Lavrentiev, C. Marchello-Nizia and T. Rainsford). All these editions should be published in 2014.

References

1. **Marchello-Nizia, Ch. and Lavrentiev, A. (ed.)** (2009-2013). *La queste del saint Graal*. Édition numérique interactive du manuscrit Lyon (BM P.A. 77). Lyon: ENS de Lyon. txm.bfm-corpus.org/txm (accessed 31 October 2013).
2. **Guillot, C., Lavrentiev, A., Rainsford, T., Marchello-Nizia, C., Heiden, S.** (2013). *La "philologie numérique": tentative de définition d'un nouvel objet éditorial*. Lemaréchal, A., Koch, P., Swiggers, P. (ed.). Actes du XXVIIe Congrès international de linguistique et de philologie romanes (Nancy, 15-20 juillet 2013). Section 13: Philologie textuelle et éditoriale. Nancy: ATILF. halshs.archives-ouvertes.fr/halshs-00846767
3. **Serge Heiden**, (2010). *The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme*. Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation, Institute for Digital Enhancement of Co
4. **Pignatelli, C., Lavrentiev, A.** (2013). *Le Psautier d'Arundel : une nouvelle édition*. Lemaréchal, A., Koch, P., Swiggers, P. (ed.). Actes du XXVIIe Congrès international de linguistique et de philologie romanes (Nancy, 15-20 juillet 2013). Section 13: Philologie textuelle et éditoriale. Nancy: ATILF. halshs.archives-ouvertes.fr/halshs-00846770

Modèles tridimensionnels pour la représentation de l'état des connaissances et propositions de visualisation pour l'analyse des corpus textuels.

Leblanc, Jean-Marc

jean-marc.leblanc@u-ppec.fr

Université Paris Est Créteil (UPEC), France

Pérès, Marie

m-peres@wanadoo.fr

Université Paris Est Créteil (UPEC), France

Résumé

Alors que la réflexion sur la visualisation est au cœur de disciplines parfois émergentes, que la web véhicule un grand nombre de représentations souvent sophistiquées, introduisant de nouveaux modèles, souvent esthétiques mais qui nécessitent de nouveaux apprentissages pour la lecture et l'interprétation, de nombreuses disciplines s'appuient actuellement sur la visualisation pour communiquer leurs résultats ou créer des outils scientifiques.

Cette contribution prend appui sur des champs disciplinaires différents pour proposer une réflexion sur la représentation des connaissances qui nous a conduit à développer un outil de visualisation des données textuelles.

Nous analyserons ici les méthodes et principes de restitution archéologiques depuis les envois de Rome jusqu'aux modèles informatiques tridimensionnels interrogeant leur forme, leur propos et le rôle de transformateur d'information que doit endosser tout créateur de restitution archéologique afin de créer un objet répondant aux contraintes de la médiatisation et la médiation de l'objet d'étude. Nous montrerons en quoi ces réflexions ont nourri la conception de l'outil TextObserver qui s'inscrit dans la lignée de la textométrie en y introduisant de nouveaux modèles.

Les outils informatiques utilisés dans le domaine de la textométrie ont longtemps fonctionné sur des modèles bien éprouvés mais reposant quasi-essentiellement sur des visuels statistiques offrant peu de possibilités de manipulation ou d'expérimentation, et parfois peu aboutis sur le plan de l'ergonomie. Face à ce constat, le logiciel développé dans le prolongement de cette réflexion, TextObserver, vise précisément à introduire de nouveaux modèles de représentation des données et des résultats pour l'analyse des corpus textuels et multimodaux. Il propose des fonctionnalités originales sur le plan de la visualisation, rendues explicites par l'interactivité, et le traitement dynamique des données et des résultats textométriques. Il rend possible l'intégration de données textuelles diversifiées dans un cadre multimédia et répond en temps réel aux questionnements expérimentaux comme les facteurs de la variation discursive.

La rencontre entre ces questionnements novateurs en lexicométrie et ceux développés avec la création du modèle archéologique tridimensionnel servent de fondement à la création de cet outil exploratoire. Nous proposons de présenter les étapes de cette réflexion puis d'exposer les fonctionnalités essentielles de TextObserver en explorant un corpus de discours politiques.

Dans un premier temps nous présenterons une réflexion sur la représentation en archéologie et médiation de l'objet d'étude. Nous reviendrons sur l'évolution des représentations, depuis la renaissance jusqu'aux envois de Rome en passant par les premières représentations scientifiques du 18ème siècle et les premiers graphiques, jusqu'à l'émergence de la notion d'infographie au vingtième siècle dans les années 70, puis les pratiques liées à la data science et data visualisation que nous connaissons aujourd'hui.

Après cet état de l'art nous en viendrons à la représentation archéologique en nous appuyant sur l'exemple de la

modélisation tridimensionnelle du Circus Maximus, mettant au jour le fait que le créateur d'objet multimédia devient un transformateur d'information.

La seconde partie de la contribution sera consacrée à la présentation de l'outil TextObserver dont la conception est issue d'une réflexion sur l'ergonomisation et la représentation des données textuelles et qui réinvestit les acquis de la recherche évoquée en première partie.

TextObserver est à la fois un outil de recherche et de formation à la recherche et s'inscrit dans une démarche proche de la textométrie, ajoutant à celle-ci une dimension expérimentale, multimodale et multimédia.

C'est en effet à partir de l'expertise approfondie d'un dispositif composé de logiciels longitudinaux et contrastifs (lexico3, Hyperbase, TXM) ou structurants (comme Alceste, iramuteq ou Astartext) voire de catégoriseurs (Cordial, Treetagger), ou d'analyseurs sémantiques (Tropes) que se fonde la conception de *TextObserver*. Il ne s'agit pas d'implémenter des fonctionnalités qui existeraient déjà au sein de ces logiciels mais d'apporter une réponse en termes d'ergonomie, d'interactivité et de visualisation et de développer des fonctionnalités originales. *TextObserver* a été conçu en outre pour répondre à des questions de recherche faisant intervenir la variation.

Ainsi *TextObserver* permet de visualiser les textes sous un angle différent, mais aussi de mieux appréhender les mesures mobilisées en textométrie. L'interactivité est l'innovation essentielle de ce logiciel: elle permet de saisir de visualisations complexes telle que l'analyse factorielle de correspondances.

Nous présenterons tout d'abord le principe de la démarche textométrique dans laquelle s'inscrit *TextObserver*, présenterons le corpus que nous prenons ici comme matériau d'expérimentation, puis montrerons au moyen d'expertises ciblées, en quoi *TextObserver* permet de mettre au jour des phénomènes de variation qui seraient difficilement appréhendables au moyen des outils logiciels classiques. Nous articulerons cette présentation autour de trois axes: les fonctionnalités de visualisation, les fonctionnalités de calcul, les fonctionnalités de navigation.

Enfin nous évoquerons les développements futurs de *TextObserver* et les premières analyses menées au moyen de cet outil sur des corpus multimodaux (donc non exclusivement textuels) qui permettent d'étendre la recherche à l'analyse du web (recueil, constitution et visualisation de corpus en temps réel, analyse automatisée de pagds web). C'est donc un élargissement de la textométrie vers une textométrie multimédia - nous justifierons cette terminologie - que nous nous proposons de présenter dans cette contribution.

References

Barats, C., Fiala, P., Leblanc, JM (2013) *Approches textométriques du web : corpus et outils*. In Manuel d'analyse du Web (Dir Christine Barats), Armand Colin, Paris, 100-124.

Benzécri J. P. (1980) *Pratique de l'analyse des données*, Dunod, Paris.

Benzécri J. P. (1982) *Histoire et préhistoire de l'analyse des données*, Dunod, Paris.

Bonin S. (1983). *Initiation à la graphique : transcription visuelle des données statistiques et cartographiques*, Épi éditeurs

Bouroche J-M., G. Saporta, (1980) *L'analyse des données*, PUF, (coll. "Que sais-je", n°1854, Paris.

Cibois P., (2000) *l'analyse factorielle*, Presses Universitaires de France - PUF (Que sais-je ?), (5^e éd.), Paris.

Daknou A. (2011) *Architecture distribuée à base d'agents pour optimiser la prise en charge des patients dans les services d'urgence en milieu hospitalier*, Thèse de doctorat, Ecole Centrale de Lille.

Gambette, P., Nuria, G., Guénoche, A., Nasr, A. (2012), *Longueur de branches et arbres de mots*, inCorpus 11, 129-146.

Guilmeau-Shala S. (2011). *En quête de la couleur : publication de dessins réalisés lors de voyages d'études en Grèce*, in Bibliothèques d'atelier. Édition et enseignement

- de l'architecture, Paris 1785-1871, INHA (« Les catalogues d'exposition de l'INHA »)
- James R. Beniger et Dorothy L. Robyn.** (1978). *Quantitative graphics in statistics: A brief history*, The American Statistician, no 32, pp. 1-11
- Lebart L., Morineau A., Piron M.**, (2000) *Statistique exploratoire multidimensionnelle*, Dunod, Paris.
- Lechleiter F.** (sous la direction de Foucart B.) (2008). *Les envois de Rome des pensionnaires peintres de l'Académie de France à Rome de 1863 à 1914*, thèse de doctorat, Université Paris IV.
- LUONG X.**, (1998) *Représenter les données textuelles par les arbres* in S. Mellet (éd.), JADT 98, Nice 1998 (avec J.P. Barthélémy).
- Pérès M.** (2006). *De la modélisation à l'image virtuelle : image et réel*. Figure de l'art, vol.(11): 197-208.
- Pérès M.** (sous la direction de Golvin JC.) (2001). *Réflexion sur le modèle informatique du Cicus Maximus*. Thèse de doctorat, Université Michel de Montaigne - Bordeaux 3.
- Plantin JC.** (2013) "D'une carte à l'autre: le potentiel heuristique de la comparaison entre graphes du web et cartes géographiques". Dans Barats C.,(dir.) in Analyser le web en Sciences Humaines et Sociales, Armand Colin, Paris.
- Roxin I., Hutschmitt B., Mercier D. and Leblanc J.M.** (2007) *Web sémantique, navigation dans de grands corpus textuels* In CIDE 10, dixième colloque international sur le document numérique, 2-4 juillet 2007, Nancy, INIST.
- Viprey J.M.** (2006) *Ergonomiser la visualisation AFC dans un environnement d'exploration textuelle : une projection « géodésique »* in actes des JADT 2006, 989-1000.
- Weiss G.** (2001) *Agent orientation in software engineering*, in The Knowledge Engineering Review, vol. 16, no. 4, 349 - 373.
- Wildbur P. et Burke M.** (2001). *Le graphisme d'information, Cartes, diagrammes, interfaces et signalétiques*, Thames & Hudson
- Wooldridge M. and Jennings N.R.** (1995) *Intelligent agents: Theory and practice*, The Knowledge Engineering Review, 10(2):115-152.

Supporting "Distant Reading" for Web Archives

Lin, Jimmy

University of Maryland, United States of America

Kraus, Kari

karimkraus@gmail.com

University of Maryland, United States of America

Punzalan, Ricardo L. Punzalan

University of Maryland, United States of America

In a recent essay on the stock footage libraries amassed by Hollywood studios in the first half of the 20th century, Rick Prelinger—moving image archivist at the Internet Archive—laments that “archives often seem like a first-aid kit or a rusty tool, resources that we find reassuring but rarely use” (Prelinger 2012). Although he doesn’t single them out by name, web archives are particularly vulnerable to this charge. User studies, access statistics, page views, and other metrics have in recent years told a consistent story: web content that has been harvested and preserved by collecting institutions, universities, and other organizations often lies fallow, and like Prelinger’s rusty tool may be notable more for its latent potential than for having served any real purpose (Hockx-Yu 2013; Kamps 2013; Huurdeman et al 2013). While the reasons for neglect are myriad, this paper focuses on one: the lack of tools to support a wide range of interactions with the content. We describe initiatives underway at the University of Maryland to partially redress the problem and highlight the need for qualitative user studies.

The Internet Archive’s Wayback Machine is perhaps the best-known and most widely available tool to browse captured content. Both the Internet Archive’s main public

site and Archive-It, its subscription-based web archiving service, replicate the experience of viewing web pages on the live web, thus reifying a “close-reading” experience. First developed in the mid-1990s, the software came of age at the same time digital humanities scholars were building the first generation of web collections aimed at providing high-resolution digital facsimiles of literary and artistic works by Blake, Rossetti, Dickinson, Whitman, and others. The emphasis on accurate rendering and display is thus a hallmark of both the Wayback Machine and many early DH projects, the latter of which likewise self-identify as “archives,” albeit archives on a dramatically smaller scale.

Although the capabilities offered by the Internet Archive and other commercial services are significant, we believe considerable technical advances are needed if web archives are to fulfill their promise as tools of analysis as well as preservation. Within the field of DH, the big data vistas offered by scholars such as Matt Jockers and Ted Underwood provide both inspiration and models on which to base these efforts (Jockers 2013). Unlike the boutique digitization initiatives that characterize the early wave of DH archives of the 1990s and early 2000s, which were often devoted to the works of a single author, the new macroanalytic approaches are premised on mass-digitization of print heritage. The paradigm they embody, moreover, is not digitization in the service of verisimilitude—reproductions that show exact fidelity to their originals—but rather digitization that produces terabytes’ worth of intermediary copies that can be cleaned, normalized, segmented, tokenized, mined, and visualized to yield new insights about the cultural record writ large. Such a paradigm disrupts the usual data-information-knowledge continuum by taking the unitary wholes of creative expression—the “cooked” novels or poems or historical documents in print—and temporarily degrading them to a “raw” data state so that they can be analyzed at scale to make higher-order knowledge claims.

We believe that the technical infrastructure to support macroanalytics or “distant reading” on web archives today is inadequate. Existing tools were built before the coming of age of “big data” technologies and provide wobbly foundations on which to build analytical tools that scale to petabytes of data. As an example, the open-source Wayback Machine is implemented as a monolithic stack primarily designed to scale “up” on more powerful servers and expensive network-attached storage. Its architecture captures the ethos of “state-of-the-art” software engineering practices of the late 1990s. Not surprisingly, the field has advanced by leaps and bounds in the last decade and a half. In the 2000s, Google published a series of seminal papers describing solutions to its data management woes, which involve analyzing, indexing, and searching untold billions of web pages. Instead of scaling “up” on more powerful individual servers, the strategy entailed scaling “out” on clusters of commodity machines (Barroso et al., 2013). Before long, open-source implementations of these Google technologies were created, bringing the same massive data analytic capabilities to “the rest of us.” These systems form the foundation of what we know as “big data” today, and provide the backbone of data analytics infrastructure at Facebook, Twitter, LinkedIn, and many other organizations. Three key systems are:

- The Hadoop Distributed File System (HDFS), which is a horizontally-scalable file system designed to store data on clusters of commodity servers in a fault-tolerant manner (Ghemawat et al. 2003). The largest known HDFS instance (by Facebook) holds over 100 petabytes.
- Hadoop MapReduce, which is a simple yet expressive programming model for distributed computations that works in concert with data stored in HDFS (Dean and Ghemawat, 2004). MapReduce models analytical tasks in two distinct phases: a “map” phase where computations applied in parallel, followed by a “reduce” phase that aggregates partial results.
- HBase, which is a distributed store for semi-structured data built on top of HDFS that allows low-latency random access to billions of records. Google’s Bigtable (Chang et al., 2006), from which HBase descended, powers Gmail, Google Maps, as well as the company’s indexing pipeline.

Modern big data technologies provide a technical path forward and an accompanying research agenda that does for web archives what macroanalytics or so-called “distant reading” has begun to do for digitized corpora in DH. As a first step in this effort, we are developing Warcbase, an open-source platform for storing, managing, and analyzing web archives built on the three technologies discussed above. The platform provides a flexible data model for organizing web content as well as metadata and extracted knowledge. We have built a prototype application that provides functionality comparable to the Wayback Machine in allowing users to browse different versions of resources in a web archive (typically as WARC or ARC files). Since Warcbase takes advantage of proven open-source technologies, we are confident of the infrastructure’s ability to scale in a seamless and cost-effective manner.

Yet Warcbase is only the beginning. We believe that our prototype—and, more generally, the technologies described above—will provide new capabilities that support innovative uses of web archives. Responsive full-text search on massive collections of web pages, one of the first items on a scholarly wishlist, is within reach: the tools exist in various open-source projects, awaiting integration. Longitudinal analyses of web pages such as tracking the frequency of person or place names become possible if we integrate off-the-shelf natural language processing tools. Yet another possibility is topic modeling on a massive scale; a separate project at the University of Maryland has built Mr.LDA, an open-source Hadoop toolkit for scalable topic modeling (Zhai et al., 2012). To provide a hint of what’s possible, we have been working with Congressional archives from the Library of Congress to explore topic modeling and large-scale visualizations of archived content, the results of which we will share during the conference presentation.

Why are large web archives so underused? It is surely not due to a lack of culturally significant material. Valuable content, ripe for exploration, ranges across topics such as electronic literature, alternate reality games, digital tools for human rights awareness, the Arab Spring uprising, and Russian parliamentary elections, to name just a few. Restrictive access regimes are partially to blame, but that alone does not provide a sufficient explanation. We believe that the issue, to a large extent, is a technological form of circular reasoning: scholars do little because the right tools don’t exist, and tool builders are hesitant to build for non-existent needs and users. Progress is necessary to understand the essential activities, methods, and questions of researchers. Interviews with current web archive users are a start, but breakthroughs will require deep collaborations between scholars and technologists. The end goal is a comprehensive set of tools for researchers in the digital humanities and beyond to analyze and explore our digital cultural heritage.

References

- Archive-It: Web Archiving Services for Libraries and Archives.** archive-it.org
- Barroso, Luiz Andre, Jimmy Clidaras, and Urs Holzle.** (2013). *The datacenter as a computer: an introduction to the design of warehouse-scale machines* (second edition). Morgan & Claypool Publishers.
- Chang, Fay, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach, Michael Burrows, Tushar Chandra, Andrew Fikes, and Robert Gruber.** (2006). “*Bigtable: A distributed storage system for structured data*.” Proceedings of the 7th USENIX Symposium on Operating System Design and Implementation (OSDI).
- Dean, Jeffrey and Sanjay Ghemawat.** (2004). “*MapReduce: Simplified data processing on large clusters*.” Proceedings of the 6th USENIX Symposium on Operating System Design and Implementation (OSDI).
- Ghemawat, Sanjay, Howard Gobioff, and Shun-Tak Leung.** (2003). “*The Google File System*.” Proceedings of the 19th ACM Symposium on Operating Systems Principles (SOSP).
- Hockx-Yu, Helen.** (15 February 2013) “*Scholarly use of web archives*.” files.dnb.de/nestor/veranstaltungen/2013-02-27-scholarly-use-of-web-archives_public.pdf

Huurdeeman, Hugo, et al. (2013). “*Sprint methods for web archive research*.” WebSci 2013 Proceedings of the 5th Annual ACM Web Science Conference:182-190.

Jockers, Matthew. (2013). *Macroanalysis: digital methods and literary history*. Urbana-Champaign: University of Illinois P.

Kamps, Jaap. (1 August 2013). “*When search becomes research and research becomes search*.” SIGIR’13 Workshop on Exploration, Navigation and Retrieval of Information in Cultural Heritage (ENRICH). Dublin, Ireland. www.slideshare.net/jaap.kamps/sigir-workshop-enrich13

Moretti, Franco. (2013). *Distant reading*. London: Verso.

Moretti, Franco. (2007). *Graphs, maps, trees: Abstract models for literary history*. London; New York: Verso.

Prelinger, Rick. (2012). “*Driving through Bunker Hill*.”

In Kraus, K. and Levi, A. (Eds.). *Rough Cuts: Media and Design in Process*. MediaCommons: The New Everyday. mediacommons.futureofthebook.org/tne/pieces/driving-through-bunker-hill

Zhai, Ke, Jordan Boyd-Graber, Nima Asadi, and Mohamad Alkhouja. (2012). “*Mr. LDA: A flexible large scale topic modeling package using variational inference in MapReduce*.” Proceedings of the 21th International World Wide Web Conference (WWW).

The term “distant reading” was coined by **Franco Moretti** in *Graphs, Maps, Trees* (2007) and has undergone further elaboration in his newest book (2013). See the “Works Cited” section for full bibliographic information.

Developing a Physical Interactive Space for Innovative Digital Humanities Exhibition

Liu, Jyi-Shane

Natioanl Chengchi University, Taiwan, Republic of China

Liao, Wen-Hung

Natioanl Chengchi University, Taiwan, Republic of China

1. Introduction

Digital humanities empower a creative transformation in both humanities and computing research by inspiring and fostering interdisciplinary interaction. Recently, digital visualization has been considered and established as a scholar methodology for digital humanities (Jessop, 2008). Projects, such as “Tooling Up for Digital Humanities” and “The Spatial History” (White, 2010) at Stanford University, have explored and experimented with various forms of graphic representation of data. Visualization is insightfully considered as part of a research process that may induce powerful arguments or raise new questions. It is also pointed out that visualization seems to give a sense of objective and scientific communication in the scholarly, yet sometime ambiguous, activities of digital humanities.

One of the less addressed issues in digital humanities visualization concerns the exhibition facilities. Even though some display equipment and technologies have been developed for some times, their innovative integration with a large-scale auditorium space to create an exhibition facilities for digital humanities has actually been little reported. We developed an innovative exhibition facility for digital humanities visualization with a conceptual framework of place-making that exploits digital technological mediation of people and humanities. Similar to the museum experiences with innovative engagement (Falk & Dierking, 2000) (McCarthy & Ciolfi, 2008), the exhibition facility induces locative experience for sense-making and potentially plays a pivotal role in facilitating further advance of digital humanities. Our work provides a field tested contribution to the research community by engaging wider audience for digital humanities, facilitating its social impact, and filling the vacancy of building a physical platform for presenting and showcasing research results for better recognition.

2. Physical Interactive Space as Digital Humanities Exhibition Facilities

Following the notion of place-making in urban development and heritage studies (Malpas, 2008), a physical space forms an existential ground where people's senses of digital humanities are shaped and defined. Therefore, an innovative exhibition facility can serve as a social and technical infrastructure of cross-disciplinary interaction and allow for new experiences with tangible and intangible forms of digital humanities. This opens up new ways of exploring and articulating digital humanities visualization with physical and social settings, and potentially widening appreciation and deepening recognition of digital humanities for general audience.

We developed the exhibition facility by transforming a large room used for library reference service and installing an array of display equipment for various forms of interactive visualization. With a floor space of 810 square meters, the room was re-conceptualized as a mixture of digital gallery and auditorium by novel interior design and technology embedment. Figure 1 shows the floor plan of the exhibition facility that comprises an inner conference room, a flanked outer corridor, and a lobby. The inner and outer space are separated by sliding doors in the front opening, auxiliary doors in the corners, and entrance doors from the lobby.

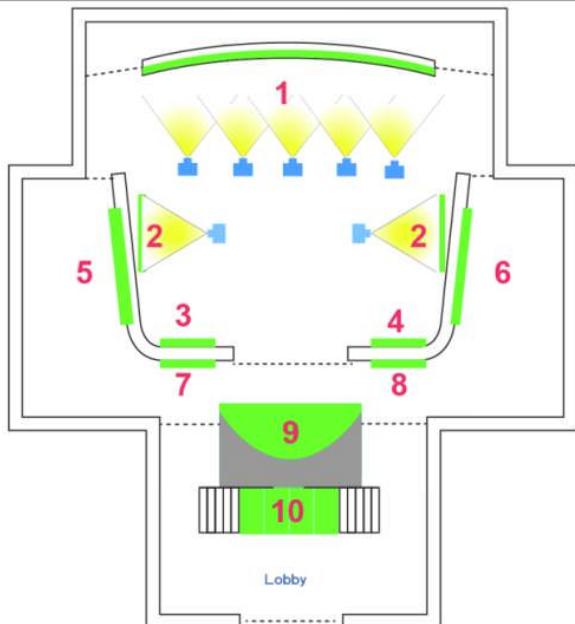


Fig. 1: Floor Plan of the Exhibition Facility for Digital Humanities

A number of ten display systems are either mounted or projected on walls in both parts of the facility, as listed below.

1. An arc wall in size of 12 meters by 2.5 meters (width and height) used as a touch wall display with projection blending of 5 projectors, rendering a surrounding effect of visualization.
2. Two 120-inch retractable projection screens, providing auxiliary displays.
3. Two 42-inch touch screens embedded in a wall book shelf, collaging digital and physical archival exhibition.
4. Two 12-inch monitors mounted on a photo collage wall, blending digital and physical image display.
5. A rear projection touch screen in size of 5 meters by 1.2 meters with projection blending of 3 projectors inside the partition wall, providing easy access and playful social interaction with digital images.
6. Two 60-inch 3D touch screens embedded in a partition wall, rendering 3D images of objects with 3D goggles.
7. A 55-inch touch screen embedded in a partition wall,
8. Two 42-inch transparent LCD boxes, exhibiting physical objects/materials inside the boxes while displaying digital information on the transparent screens.
9. A curvier arc wall in size of 8 meters by 2.5 meters used as a surrounded wall display with projection blending of 3 projectors, rendering immersive visualization.
10. A collage of wall-mounted four 46-inch screens in 4K2K resolution (ultra high definition), used as a digital signage board in the lobby.

Figure 2 through Figure 5 show actual images of the renovated results for an innovative exhibition facility.



Fig. 2: Evacuating a room previously used for library reference service



Fig. 3: Renovated as a conference room and auditorium, showing display systems #3 and #9 in Figure 1



Fig. 4: Part of the corridor flanking the inner room, showing display systems #5 and #7 in Figure 1



Fig. 5: Renovated lobby, showing display system #10 in Figure 1.

Figure 6 through Figure 9 show some of the exhibition highlights from a range of



Fig. 6: A workshop for digital humanities visualization in the conference room



Fig. 7: An international visitor appreciating an ancient book inside the transparent box, while getting information on the touch screen



Fig. 8: A group of students enthusiastically interacting with a large scale touch screen



Fig. 9: A group of international visitors enjoying a 3D digital simulation of the cultural heritage of lantern festival

3. Digital Presentation and Exhibition of Digital Humanities

The developed facility provides intensive and large scale visualization in an atmosphere with aesthetics appeal (Guyer, 2004). Large sized interactive touch screens facilitates audience engagement and creates more persuasive communication. The integration of space and technology in the exhibition facility aims to create a sense of place with a prominent context of digital humanities in which a living and sustainable recognition with exhibited subjects can be induced. Exhibition audience is, therefore, contextualized with senses, feeling, and embodiment that underpins an interpretive process of meaning-making (Schorch, 2012) and leads to an internal understanding and empathy of digital humanities.

The exhibition facility has been completed and inaugurated in May 2013 and has offered a range of subjects on a regular basis. The place becomes a hot spot for campus activities and has been designated to receive dignified visitors. Audience generally expresses experiences of remarkable novelty and deep appreciation for digital humanities exhibition. The innovative facility transformation project has been regarded as an overwhelming success by both inside and outside of campus community and has strengthen the initiative of digital humanities as a university research agenda.



Fig. 10: Geographic distribution of analytic data being projected on the 12-meter arc wall.

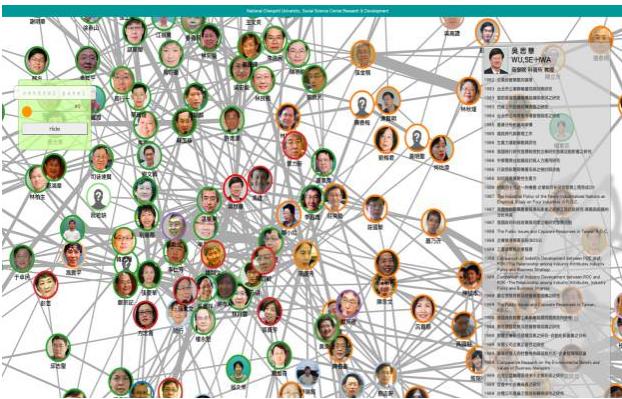


Fig. 11: Cooperative research networks among faculty members with interactive query on touch screen

We illustrate two use cases of the facility. The first is for visualization support of analytic investigation. Figure 10 and Figure 11 show images of analytic data used in research meeting of digital humanities projects. It has been indicated that large scale visualization of exploratory data inspection process achieves effective communication and facilitates research progress. An interactive script of images also helps present research discovery to the general audience.



Fig. 12: The author digitally interacts with his own handwritten manuscripts, along with his pupil.



Fig. 13: Manuscripts of short articles are collaged on a wall with digital images of places in the articles.

The second use case is an honorary ceremony for university chair professor, also a revered writer, along with an exhibition of his highly regarded books and original handwritten manuscripts over thirty years. Figure 12 through Figure 15 show a range of presentation forms to provide a rich context for the writer's celebrated career. The chair professor was apparently moved by the immersive atmosphere that reflected his heartfelt memory of life.



Fig. 14: The chair professor gave a talk to many of his students and readers in the conference room



Fig. 15: Students listen to a poem recital in author's recorded voice with an immersive textual and graphic background.

4. Conclusion and Future Work

We conclude that an innovative exhibition facility is a vital infrastructure for digital humanities endeavor. It is argued that the existential grounding of digital humanities is better achieved with an engaging and interpretive process in an immersive atmosphere. This meaning-making system helps

gather enthusiastic support for and arouses great interest in digital humanities. More importantly, the facility provides a playground for digital humanities activities and establishes a digital laboratory for digital humanities research.

Another implication is the exciting potential in pedagogical use. With project-based learning, interdisciplinary student teams are taught to digitally curate subjects of cultural heritage and to organize innovative exhibitions for their peers. Ongoing student projects include Chinese puppet show and a local heritage village with colorful house wall painting. We believe that the educational value of the exhibition facility will contribute to wider participation of younger generation in digital humanities.

We also remark that the exhibition facility along with the activities and presentation content that it facilitates are not applicable to a controlled comparison in a research lab. However, since its inauguration, the facility has hosted more than 100 school-level events with 2500 highly-impressed visitors in six months. We feel that this actual usage result provides stronger verification than lab experimental data. A preliminary sampled survey showed significant effects on enabling audience recognition and appreciation of digital humanities activities. Our future work includes a more formal user study.

References

- Falk, J. H., and Dierking, L. D. (2000). *Learning from Museums: Visitor Experiences and the Making of Meaning*. Altamira Press.
- Guyer, P. (2004). *The Origins of Modern Aesthetics: 1711-35*. In I. Kivy (Ed.), *The Blackwell Guide to Aesthetics*. Oxford, UK: Blackwell Publishing.
- Jessop, M. (2008). *Digital Visualization as a Scholarly Activity*. *Literary and Linguistic Computing*, 23(3): 281-193.
- McCarthy, J. and Ciolfi, L. (2008). *Place as Dialogue: Understanding and Supporting the Museum Experience*. *International Journal of Heritage Studies*, 14(3), pp. 247-267.
- Schorch, P. (2012). *Cultural Feelings and the Making of Meaning*. *International Journal of Heritage Studies*, pp. 1-14.
- White, R. (2010). *What Is Spatial History?*, <http://www.stanford.edu/group/spatialhistory/cgi-bin/site/pub.php?id=29>.

Mining the Cloud of Witness: Inferring the Prestige of Saints from Medieval Paintings

Lombardi, Thomas
 tlombardi@washjeff.edu
 Washington & Jefferson College

1. Introduction

1.1. Overview

Previous research has demonstrated the utility of constructing undirected, weighted networks from the co-occurrence of people in images.¹ Researchers have repurposed the technique to analyze the evolution of iconography in medieval artwork.² Using this technique, when two saints appear together in an image, the nodes representing these saints are linked. Moreover, each time these saints appear together the weight on the link connecting them increases. This technique captures the evolution of these co-occurrences in revealing ways (Fig. 1).³ For example, lines of high weight capture important motifs in the artwork such as the links between Christ and Mary based on the common Madonna and Child image. In a corpus of early images of Saint

Francis, the evolution of this network captured the development of a stable core of saints consistent with the historical and artistic record.

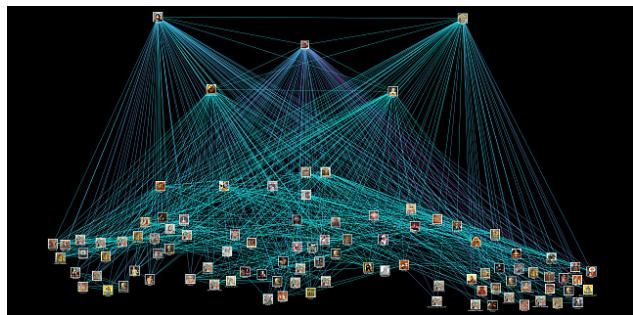


Fig. 1: Undirected weighted network of saints created in Cytoscape

Unfortunately, these networks do not capture the hierarchies in such iconography. For example, the presence of Saint Anthony of Padua, a Franciscan saint, depends on the presence of Francis. Historically, Francis was the head of the Franciscan order and therefore Anthony depended on the institution bearing Francis' name. Artistically, Anthony does not appear without Francis in imagery at this time.⁴ In order to capture the development of this imagery, including the hierarchical aspects of the organization of these saints, the technique must capture the direction of links representing these dependencies. This paper demonstrates that association rule mining allows for the creation of directed, weighted networks of saints. The social prestige of the saints can then be inferred from the structural prestige observed in the network.

1.2. Methodology

The corpus includes 236 images of Saint Francis of Italian production from 1230 to circa 1320 and serves as an excellent case study for this technique.⁵ The catalog provides information about dating, provenance, authenticity, style and documentation. The iconography of Francis provides a dramatic example of a transition from regionally-venerated to internationally-venerated saint. Given this complex transition, the resulting network captures interesting shifts in the thematic content of the iconography. The proposed technique along with previously developed network models provide powerful ways to explore iconographic trends.

Constructing the directed network involves standard techniques in data mining⁶ and network analysis⁷. Cook's corpus was converted into a matrix for rule mining in *RapidMiner Studio*.⁸ The denormalized data captures the presence or absence of a saint in each painting. The resulting matrix includes 236 rows representing the paintings and 102 columns representing saints. After preparing the data, support and confidence metrics for each pair of saints was calculated. The confidence metrics model the strength of the relationship between saints in each direction, providing a basis for inferring rank in the relationships (Fig. 2). For example, the thick pink link from Anthony to Francis represents a confidence of 1.0, meaning that every time Anthony appears in an image Francis certainly appears as well. The thin blue link from Francis to Anthony, on the other hand, has a confidence of 0.122, signifying that Francis appears without Anthony in many paintings. The difference between these confidence metrics determines the weight and direction of the link between the two saints. In this example, we would replace the two links pictured with a single link from Anthony to Francis bearing a weight of 0.878. With this directed network, we can calculate the structural prestige of the saints in *Pajek*.^{9 10}

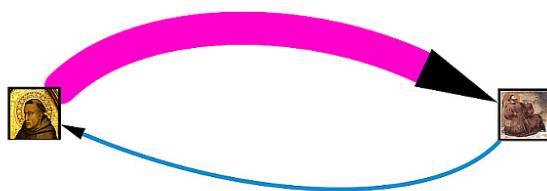


Fig. 2: Directed weighted links between Anthony of Padua (left) and Francis

For the purposes of demonstration, the calculations have been performed on a small set of data. Table 1 shows the weight calculations based on the confidence measures derived from three images. When these links are combined (Fig. 3), they produce a directed, weighted network well-suited to determining prestige. Directed networks provide several straight-forward techniques for calculating structural prestige.¹¹

The input degree is the number of links pointing to a node. In the sample network, Francis has an input degree of 3 while Gregory IX has an input degree of 0. Although input degree is often illuminating, this measure of prestige only addresses direct connections. The input proximity prestige, on the other hand, uses both direct and indirect links in its calculations of popularity. Table 2 summarizes the calculations required to compute the input proximity prestige in a directed network. Each of these metrics highlights Francis as the most prestigious saint in this simple network.

Antecedent->Consequent	Confidence(A->B)	Confidence(B->A)	Link Weight
Seraph-->Francis	1.0	0.3333	0.6667
Gregory IX-->Francis	1.0	0.3333	0.6667
Narni-->Francis	1.0	0.3333	0.6667
Seraph<--->Narni	1.0	1.0	1.0

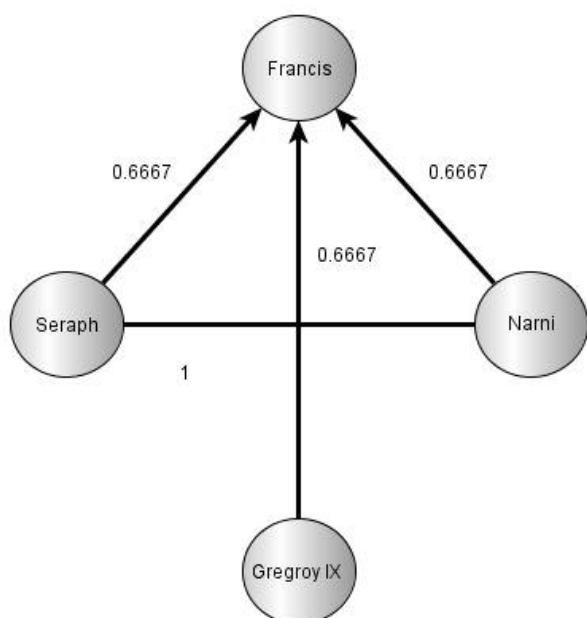


Fig. 3: Directed weighted network of saints

Saint	Influence Domain	Proportional Distance	Average Distance	Proximity Prestige
Francis	3	1.00	1.00	1.00
Seraph	1	0.33	1.00	0.33
BartholomewI of Narni		0.33	1.00	0.33
Gregory IX	0	0.00	Undefined	0.00

2. Results

With the additional interpretive power of directed networks, researchers can better understand changes in the popularity of saints. Somewhat surprisingly, the painters and patrons of the earliest surviving images of Francis (1230-1249) did not seek to juxtapose Francis with Christ, Mary or other well-known saints. Instead, the early promoters of Franciscan iconography chose to portray Francis by himself or with other prominent figures in Francis' hagiography such as the Seraph. As the cult of Francis grew, however, the prestige of Christ and Mary jumped from nil in the 1240s to 0.97 and 0.79 respectively in the 1320s. Moreover, by examining the correlation of prestige measures of different saints, researchers can identify trends in the artistic tastes of the patrons driving the demand for these images. For example, the prestige of the Seraph is negatively correlated with the prestige of Clare (-0.82), Benedict (-0.5), and Dominic (-0.6) meaning that when the Seraph is popular in this imagery the monastic and mendicant leaders are not. This negative correlation echoes Cook's observations regarding the presentation of Francis in non-Franciscan houses, particularly that Francis is often presented without stigmata in this context.¹²

Apparently, non-Franciscan houses and even the Clares did not wish to emphasize the unique aspects of Francis' hagiography in their commissions. The prestige figures also register the effects of specific events such as the canonization of saints. For example, Louis of Toulouse was canonized on April 7th, 1317; his prestige in the decade 1310 to 1319 rose from nil to 0.33. During this same decade, the prestige of Anthony of Padua, another male Franciscan saint, plummeted to 0.0 from 0.45. As a popular new saint, Louis displaced Anthony for about a decade as the preferred male Franciscan to balance compositions with Francis. Finally, the prestige metrics indicate a growing popularity of female saints after 1300. Excluding Christ, 4 of the 5 most popular saints in these images are female: Mary (0.79), Clare (0.49), Mary Magdalene (0.45) and Catherine of Alexandria (0.42).

Given the results of this study, we aim to expand the research to include a wider range of images. In particular, the inclusion of works of Dominican provenance will be helpful for comparing prestige metrics in both mendicant traditions.

Beyond this, we believe that applying the technique more broadly may shed light on some of the larger trends and issues in medieval art history. For example, researchers have noted that new saints and new imagery related to saints appeared in medieval art in response to the Black Death.¹³ With this technique, we can gauge the relative prestige of saints before and after the Black Death to determine if the perception of these saints changed in the eyes of painters and their patrons.

References

1. Golder, S. (2008). *Measuring Social networks with digital photograph collections*. In Proceedings of the 19th ACM Conference on Hypertext and Hypermedia, 43-48.
2. Lombardi, T. (2013). *The Communion of the Saints: Networks and the Study of Iconography*. Presented as a contributed talk at Arts, Humanities, and Complex Networks – 4th Leonardo Satellite Symposium at NetSci2013. artshumanities.netsci2013.net
3. Smoot, M., Ono, K., Ruscheinski, J., Wang, P.-L., & Ideker, T. (2011). *Cytoscape 2.8: new features for data integration and network visualization*. Bioinformatics 27(3): 431-432. Homepage: www.cytoscape.org/

4. Cook, W. R. (1999). *Images of Saint Francis in Painting, Stone and Glass from the Earliest Images to ca. 1320 in Italy: A Catalogue*. Italian Medieval and Renaissance Studies 7. Leo S. Olschki.
5. Cook, W. R. (1999). *Images of Saint Francis in Painting, Stone and Glass from the Earliest Images to ca. 1320 in Italy: A Catalogue*. Italian Medieval and Renaissance Studies 7. Leo S. Olschki.
6. Agrawal, R., Imieliński, T. & Swami, A. (1993). *Mining association rules between sets of items in large databases*. Proceedings of the 1993 ACM SIGMOD international conference on Management of data - SIGMOD '93. p. 207.
7. De Nooy, W., Mrvar, A. & Batagelj, V. (2011). *Exploratory Social Network Analysis with Pajek*. 2nd Edition. Cambridge: Cambridge University Press.
8. RapidMiner Studio. Home page: rapidminer.com/products-2/rapidminer-studio/
9. Mrvar, A. & Batagelj, V. *Pajek - Program for Large Network Analysis*. Home page: pajek.imfm.si
10. De Nooy, W., Mrvar, A. & Batagelj, V. (2011). *Exploratory Social Network Analysis with Pajek*. 2nd Edition. Cambridge: Cambridge University Press.
11. De Nooy, W., Mrvar, A. & Batagelj, V. (2011). *Exploratory Social Network Analysis with Pajek*. 2nd Edition. Cambridge: Cambridge University Press.
12. Cook, W. R. (1999). *Images of Saint Francis in Painting, Stone and Glass from the Earliest Images to ca. 1320 in Italy: A Catalogue*. Italian Medieval and Renaissance Studies 7. Leo S. Olschki, p. 103.
13. Meiss, M. (1951). *Painting in Florence and Siena after the Black Death*. Princeton: Princeton University Press.

Detection of Poetic Content in Historic Newspapers through Image Analysis

Lorang, Elizabeth M

University of Nebraska-Lincoln, United States of America

Soh, Leen-Kiat

University of Nebraska-Lincoln, United States of America

Lunde, Joseph

University of Nebraska-Lincoln, United States of America

Thomas, Grace

University of Nebraska-Lincoln, United States of America

By conservative estimates, several hundred thousand poems appeared in early American and U.S. newspapers from the eighteenth through the early twentieth centuries. Counting snippets of verse that appeared in death notices, advertisements, and articles makes the presence of poetry in historic newspapers even more pervasive. Feminist scholars and others performing recovery work routinely resurrect authors and works from newspaper pages, but until recently this rich trove of newspaper verse as a corpus of its own has been outside the scope of literary study and a footnote in histories of American newspapers. In the last decade, however, scholars have made significant inroads in studying the importance of newspaper verse as a form and the public role of poetry in American culture. Underpinning this scholarship is a growing recognition that the evaluation and history of American poetry should not be based on less than one percent of the poetic record. In addition, this new scholarship values and explores the role of poetry in the daily lives of people, including making sense of what it means to be human and in processing national, social, and individual experiences. To the extent that these new histories depend on traditional methods of archival discovery and analysis, however, they will remain anecdotal—individual narratives extrapolated from a minuscule subset of the whole, with limited means of situating the anecdote as either representative or idiosyncratic. In short, the magnitude of the corpus requires new modes of discovery and analysis.

A fundamental problem in the reappraisal of newspaper verse has been finding and processing poetic content in an efficient manner, which is essential for developing new interpretations, analyses, and literary histories. The primary means of finding this content typically involves paging through original issues of newspapers, scrolling through reels of microfilm, and browsing digital images to scan, by human eye, each page for graphical features that resemble poetry. Dealing with only daily newspapers for a single year, 1860, would require visually scanning nearly half a million newspaper pages. Certainly no individual in a lifetime could complete a count—to say nothing of a comprehensive bibliography or macro-level analysis—of newspaper verse using this strategy.

While the digitization of historic newspapers has mitigated some issues of access, the main avenues for discovery in these collections are browsing and text-based searching. Browsing for poetic content in such collections follows the strategy outlined above: going image by image through digitized pages and visually scanning the images for features typical of printed poetry. Ironically, web interfaces and variations in Internet connection speeds can make digitally paging through a newspaper a slower process than either scrolling through microfilm or flipping physical pages.

How, then, might one discover poetic content in digitized historic newspapers? Our research shows that digital page images hold significant promise for scholarly inquiry with regard to poetic content. The basis for our approach is that the appearance of poetic content usually follows certain patterns that can be visually differentiated from other published texts in newspapers. Given a newspaper page, a person can survey the page and figure out quickly whether it contains a poem, to a certain degree of accuracy, without having to read or understand the text. Our project has the computer do the same visual processing as the human eye and brain when a person moves through a newspaper issue looking for poetic content. This image processing approach can also be used as a powerful filter, removing materials from further consideration that do not meet the specified criteria. That is, not only does the process work to identify pages that appear to include poetry, but it discards those that do not, weeding out much of the noise.

Methodology

The image processing component of the project consisted of two important phases: training and deployment. During the training phase, the goal was to produce a classifier able to categorize an image as either a poem or non-poem image. To produce the classifier, a training dataset was prepared and fed into a machine learning-based classifier. After the classifier was produced, we moved to the deployment phase. During this phase, the steps used to prepare the training sets were streamlined to automatically process and classify new images. In this paper, we focus on the four stages of the training phase: (1) pre-processing, (2) feature extraction, (2) neural networks learning, and (4) testing.

The first stage involved manual extraction of image snippets from digitized newspapers. For training, we developed three sets of snippets: (1) at least part of a poem appears in the snippet, (2) the snippet contains no poetic content, and (3) the snippet has visual cues that are similar to poetic content. Because each snippet is inherently noisy and could be of low quality, we performed 3x3, 5x5, and 7x7 averaging to smooth out noisy pixels, a step known as blurring. To convert the blurred snippet into a binary image—effectively identifying the object pixels from background pixels—we used a bi-Gaussian (or bi-normal) curve approximation (Haverkamp et al., 1995) to obtain the binary segmentation threshold. (See Figure 1 and Figure 2.)

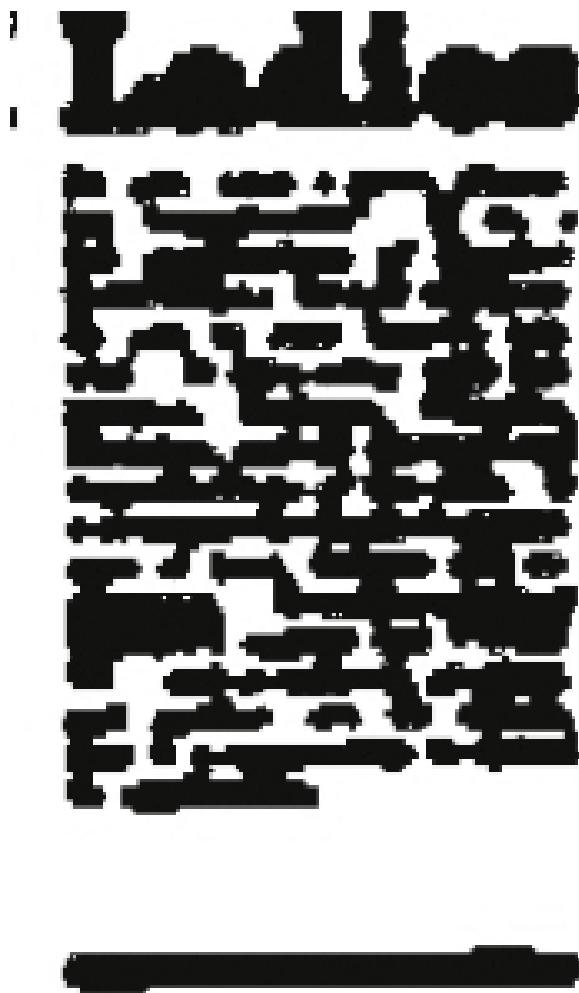


Fig. 1: Binary non-poem image snippet; 5x5 blurring.



Fig. 2: Binary poem image snippet; 7x7 blurring.

After obtaining binary images, the next task was representing and extracting visual cues as salient features. We evaluated three sets of attributes: (1) the left and right margins (number of columns without a dark pixel); (2) the vertical white spacing between adjacent lines of text (mean and standard deviation); (3) the jaggedness of the ends-of-lines for poetic content (mean and standard deviation).

Then, with this imagery data re-represented as numeric data, we used machine learning techniques to train a classifier. Specifically, we used the machine learning approach known as artificial neural networks (ANNs) (Hopfield, 1988; Yegnanarayana, 2009). An ANN learns a vector of weights on features in the dataset to choose labels for new data. Such a network consists of multiple nodes connected to threshold functions or additional layers of nodes. The network is updated iteratively until it correctly predicts labels for training data. We used a back-propagation ANN (Hecht-Nielsen, 1989) where weights connecting the nodes update iteratively based on how the network classifies known data points in the training dataset —whether its classification matches the labels of training data points.

Back-propagation ANNs have two phases. During phase 1, a training instance or data point is fed into the ANN's multi-layer structure, generating activations in the nodes and resulting in a final output label. During phase 2, "rewards" are propagated back to the input layer from the output layer based on whether the final output label matches the ground-truth label of the instance. If the output label is correct, all the linkages within the structure contributing to the correct output are rewarded with an increased weight. If the output label is incorrect, all contributing linkages are penalized accordingly with a reduced weight. In this manner, the network learns incrementally to find a combination of weights for these links.

Finally, to validate the accuracy of the classifier, a ten-fold cross validation process was used. In this process, the total data set was broken into ten groups. The classifier was

trained using nine of the groups and then tested on the single remaining set. This process of training and testing was repeated until each group had been used as the test set once. The results were then aggregated and the accuracy computed.

Findings

In addition to discussing the importance of newspapers and newspaper verse for American literary history and the need for new modes of discovery in digitized collections, this paper will report on three results of our research: (1) basic analysis, (2) training and classification analysis, and (3) a comparative study. The basic analysis will present the algorithms we used to extract visual features from the snippets, and the correlation analyses we did to ascertain the feasibility of our image-based approach. This will show the potential discrimination power of our visual features in distinguishing poem or non-poem snippets. The training and classification analysis will document our experimentation with different ANN configurations and report on our training processes. This will include the number of hidden nodes in our ANNs used, learning weights and momentum parameters investigated, convergence rates of the different configurations, and training and classification accuracies. Finally, we will report on our comparative study in terms of the usefulness of our blurring processes and how better to fuse them. In particular, we will show whether submitting each image as a single data point with three sets of attributes or submitting each blurred image with its set of attributes as a single data point leads to more effective training and higher classification accuracy.

References

- See, for example, **Bennett** (2003); **McGill** (2003); **Barrett** (2012); **Barrett and Miller**,(2005); **Gardner** (2009); **Cohen** (2010); **Garvey** (2012); **Chasar** (2012); **Rubin** (2007). In his foundational history of American journalism, **Frank Luther Mott** estimated the number of U.S. newspapers in existence in 1860 at 3,000, 11 percent of which were dailies. Most dailies in this period were four pages long.
- Barrett, F. and Miller, C.** (2005). *Words for the Hour: A New Anthology of American Civil War Poetry*. Amherst, University of Massachusetts Press.
- Bennett, P.B.** (2003). *Poets in the Public Sphere: The Emancipatory Project of American Women's Poetry, 1800-1900*. Princeton, Princeton University Press.
- Chasar, M.** (2012). *Everyday Reading: Poetry and Popular Culture in Modern America*. New York, Columbia University Press.
- Cohen, M.** (2010). *Contraband Singing: Poems and Songs in Circulation During the Civil War*, *American Literature*, 82(2): 271-304.
- Gardner, E.** (2009). *Unexpected Places: Relocating Nineteenth-Century African American Literature*. Jackson, University Press of Mississippi.
- Garvey, E. G.** (2012). *Writing with Scissors: American Scrapbooks from the Civil War to the Harlem Renaissance*. Oxford, Oxford University Press.
- Haverkamp, D., L.-K. Soh, and C. Tsatsoulis.** (1995). *A Comprehensive, Automated Approach to Determining Sea Ice Thickness from SAR Data*, *IEEE Transactions on Geoscience and Remote Sensing*, 33(1): 46-57.
- Hecht-Nielsen, R.** (1989). *Theory of the Backpropagation Neural Network*, Proceedings of the International Joint Conference on Neural Network (IJCNN'1989), Washington, DC, June 1989.
- Hopfield, J. J.** (1988). *Artificial Neural Networks*, *IEEE Circuits and Devices Magazine*, 4(5): 3-10.
- McGill, M.** (2003). *American Literature and the Culture of Reprinting, 1834-1853*. Philadelphia: University of Pennsylvania Press.
- Mott, F. L.** (1942). *American Journalism: A History of Newspapers in the United States through 250 Years, 1690 to 1840*. New York, Macmillan.

Rubin, J. S. (2007). *Songs of Ourselves: The Uses of Poetry in America*. Cambridge, MA, Belknap Press.

Yegnanarayana, B. (2006). *Artificial Neural Networks*. New Delhi, Prentice-Hall of India.

Visualizing Global News

Losh, Elizabeth

University of California, San Diego, United States of America

Manovich, Lev

CUNY Graduate Center, New York, United States of America

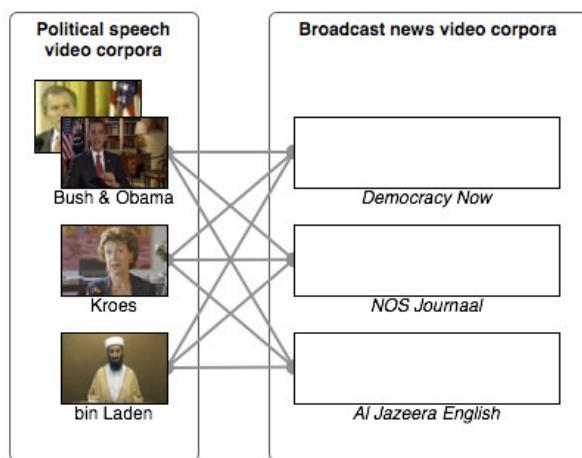


Fig. 1: Elizabeth Losh and members of Software Studies Initiative explore media visualization of 113 video public addresses by U.S. President Barack Obama. Visualizations are available at <http://lab.softwaresstudies.com/2011/09/digging-into-global-news.html>

Television news often serves as the first, most vivid draft of history and shapes the conventions of political speech and civic participation around the world, but undertaking systematic analysis of large digitized corpora of broadcast news video archives and even smaller corpora of government information videos in the public record presents a number of technical, methodological, and institutional challenges. Although a deeper understanding of how the news is represented in moving images promises to improve access to historical records, participation in public dialogue, and education at all levels, large corpora of video news programs are considerably more difficult to catalog, mine, and visualize than text news collections. Furthermore, new forms of dissemination, annotation, and commentary made possible by the Internet are rapidly transforming video news consumption on a global scale. At the same time, traditional newspapers of record are incorporating more multimedia content, so new computational methods such as video search will become increasingly necessary for researchers working with all news collections. By collaborating with archivists and computer scientists, it is possible to analyze visual rhetoric in very large video collections using media visualization

"Visualizing Global News" studies how excerpts from large collections of political speech videos with historically significant personages are remixed into even larger collections of news broadcasts. It applies new techniques in search and visualization to aid both humanities researchers and the greater public interested in exploring the visual details, aesthetic features, and narrative context of source footage that is reused in news broadcasts. The project focuses on materials from four political figures who appear in video news from 2001 to 2011 and tags and visualizes a rich set of intersecting corpora: newscasts from the US program Democracy Now!, the Qatar news channel Al Jazeera English, the top-rated Dutch news show NOS Journaal. This video corpus contains over 30,000 separate news programs, and over 2,000 videos of political speeches.

As participants in the Digital Media Analysis, Search and Management (DMASM) international workshops, we are well aware of the technical problems plaguing automatic systems and that even distinguishing foreground objects from backgrounds can be challenging. Nonetheless, the MediaMill technology has performed well in the yearly TRECVID benchmark competition, and ImagePlot can create compelling visualizations working with key frames. Our presentation aims to provide a policy overview of the opportunities and challenges of DH with global video news archives. Currently the “big data” problem of identifying heterogeneous sources in moving image archives has been largely funded by corporations interested in protecting intellectual property and state entities interested in surveillance. Without digital humanities scholars playing a larger role and generating research questions that merit publication, the most vivid aspects of the historical record are likely to remain under-theorized and analyzed without large-scale comparative study across nations and periods.



Visual rhetoric has a long tradition in the humanities that includes analysis of symbolic objects in portraits of world leaders or the choreography of their oratorical performances. However, contemporary news broadcasts present historical figures in the public record in increasingly visually complex ways that use massive collections of heterogeneous and frequently unsourced footage, motion graphics, and digital effects.

In this project we use new computational and visualization techniques to foster new forms of humanities scholarship and public access to historical records. Political figures produce memoirs, letters, editorials, and other forms of written discourse, but the speeches they compose are also performed and thus can be analyzed as much more than written texts. Now that the speeches of contemporary political leaders are recorded and archived, humanists have a rich record of public rhetoric to analyze that includes facial expression, bodily gesture, vocal performance, and frequently the use of sets and props. In the era of digital video new editing and compositing techniques are also part of the “official version” of a given speech, and portions of these speeches may be further edited and composited when they appear as part of news broadcasts. Broadcasters may adopt (or reject) signature visual styles that brand their programs and even signal their political orientation, and these markers of visual rhetoric can be mapped over time or represented comparatively (Manovich, 2011b). Scholars in the fields of rhetoric, performance studies, history, journalism, film and media studies, and civic education can compare and contrast visual and verbal political messages and changes over time in large video collections.

While a transcript of a given speech may give humanities scholars information about word use and specific references to people, places, and events, much of the rhetoric of the news is actually nonverbal. For example, the use of slow motion, freeze frame, or replay techniques may change the meaning of a given rhetorical moment. Visual arguments in news programs are now often advanced by compelling editing, dazzling information

graphics, or aesthetic choices that privilege aspects such as warm lighting or patriotic color schemes. At the same time, the size of digital collections of video news is growing rapidly in response to a number of trends: 1) cable television and satellite television, which spawned the twenty-four-hour news cycle, have become international phenomena, 2) television stations now create web-only content to draw Internet viewers to their programs online, 3) newspapers are incorporating multimedia content and interactive features in archives that once only indexed print news, and 4) bloggers and vloggers involved in citizen journalism have participated in a dramatic expansion of the scope and scale of independent media.

The difficulty of humanities news analysis is exacerbated by the fact that a typical television news program is actually “database cinema” (Manovich 2005) composed of many different kinds of source footage, including studio shots of anchors, field reporting, tock footage, video news releases, government public relations materials, and witness journalism from cell phones and mobile devices. Source clips may not be attributed much less tagged with date, location, individuals shown, or person/organization behind the camera (Losh 2008, Gregory 2010), and the meaning of objects or gestures in the frame, as evidence of the intention or motivation of particular political actors, can be controversial (Losh 2011).

The UCSD project team created visualizations with the official digital video archive from the Obama administration at WhiteHouse.gov that focused on the collection of “Weekly Addresses” directly addressing an imagined American Internet viewer that includes extensive discussions about current events that range from repeated occurrences (economic boosterism, holiday celebrations, etc.) to one-time disasters and tragedies (the Deepwater Horizon oil spill, the shootings at Fort Hood, etc.). UCSD has also examined a significant subset of videos in the WhiteHouse.gov archive that are addressed to audiences abroad as part of U.S. public diplomacy campaigns, which are sometimes also remixed into global news broadcasts. Even a country with a relatively small population or one that is not usually considered critical to U.S. interests, such as Côte d'Ivoire, may have a designated Obama direct address. This dataset allows those studying the visual rhetoric of international relations to work with a particularly rich and dense corpus that includes acknowledgement of cultural exchanges and national holidays. For example, historians studying U.S. Iranian relations in the twenty first century could look closely at three separate annual addresses recorded by Barack Obama that were intended to go directly to the people of Iran on the occasion of Nowrūz, the Persian new year. Each address is fundamentally different in tone and diction, as the U.S.-Iranian diplomatic relationship seems to deteriorate, as well as in shot composition and White House location.

Media created by a sitting U.S. president are always historically significant, but media created by Barack Obama -- the first test case area in this project -- are especially interesting to humanities researchers. In addition to scholarship done by rhetoricians, historians, political scientists, and performance studies and communication scholars, Barack Obama has also generated significant scholarly attention worldwide from those who study American popular culture, race relations, religion, gender, class, civic participation, digital culture, and globalization. The volume of scholarly publication of peer-reviewed books and articles about Obama and the large number of conferences, panels, and talks reflect the potential importance of this collection to scholarly discourse.

The Obama official video corpus has also been significant for the international press, which often uses content from the weekly addresses and other WhiteHouse.gov speeches and public statements in news broadcasts shown worldwide. Researchers at UCSD have noted the presence of material from the official Obama corpus in global broadcasts that include shows from the BBC, PressTV in Iran, Al Arabiya in the United Arab Emirates, ABS-CBN in the Philippines, and many others. The expense of maintaining news bureaus abroad is an obvious reason to rebroadcast free HD footage, but stations often add their own visuals and editorial commentary. For example, the Indian television network NDTV cropped official U.S. government footage from “President Obama’s Statement on Credit Downgrade” and added its own corporate branding along

with a skeptical digital banner that read "Obama Assures But Dow Plunges."



Fig. 3: UCSD visualization of White House government records footage of an Obama speech to the Iranian people (top) and visualizations of how the same speech appears in Democracy Now! (middle) and Al Jazeera English (bottom). To create these visualizations, the UCSD team first used open source software to automatically detect shot boundaries in the video, and then applied media visualization tools to create a grid of images where each image is the first frame of each UCSD visualization of White House government records footage of an Obama speech to the Iranian people (top) and visualizations of how the same speech appears in Democracy Now! (middle) and Al Jazeera English (bottom). To create these visualizations, the UCSD team first used open source software to automatically detect shot boundaries in the video, and then applied media visualization tools to create a grid of images where each image is the first frame of each shot.

Working with a large set of news archives from different sources will also allow us to develop possible answers to a number of critical research questions for archivists: How can we build systems that allow users to compare historical world events from different national or political perspectives? How can we connect one nation's collections to the collections of others, and how will those connected collections benefit users? Given the capabilities of new search tools, how can we best improve cataloging processes, which currently depend on costly manual input, to make collections broadly usable and accessible? How can we begin to trace the re-use and re-contextualization of primary materials, in particular, news materials?



Fig. 4: President Obama's 2009 Nowruz address to the Iranian people as excerpted by the TV news program Democracy Now! Each frame represents one shot of the program. The frames are arranged in the order of shots (left to right, top to bottom). Shot 19 (third column, third row) is the excerpt from President Obama's video address.



Fig. 5: President Obama's 2009 Nowruz address excerpted by Al Jazeera TV broadcast. Each frame represents a sequential shot of the program. Shots 5 and 6 (first row) are excerpts from President Obama's video address.

References

- American Council of Learned Societies** (2006). *Our cultural commonwealth : The report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences*. New York, NY: American Council of Learned Societies.
- Atkinson, Joshua** (2010). *Alternative media and politics of resistance: A communication perspective*. New York, NY: Peter Lang.
- Becker, Konrad** (2009). *Deep search: The politics of search beyond Google*. Innsbruck; Piscataway, N.J.: Studien Verlag.
- Boler, Megan** (2008). *Digital media and democracy tactics in hard times*. Cambridge, MA.: MIT Press.
- Borgman, Christine L.** (2010, May). *The digital archive: The data deluge arrives in the humanities*. Time Will Tell, But Epistemology Won't. In Memory of Richard Rorty. Presentation at A Celebration of Richard Rorty's Archive, University of California, Irvine.
- Dayan, Daniel** (1992). *Media events: the live broadcasting of history*. Cambridge, MA: Harvard University Press.
- Douglass, Jeremy** (2009, January). *Computer visions of computer games: Analysis and visualization of play recordings*. Workshop at Media Arts, Science, and Technology (MAST) 2009: The Future of Interactive Media. University of California, Santa Barbara.
- Douglass, Jeremy, Huber, William & Manovich, Lev** (2011). *Understanding scanlation: How to read one million fan-translated manga pages*. Image and Narrative. Brussels, Belgium.
- Edelman, Murray** (1988). *Constructing the Political Spectacle*. Chicago: University Of Chicago Press.
- Entman, Robert** (1989). *Democracy without citizens: Media and the decay of American politics*. New York, NY: Oxford University Press.
- Etzioni, A.** (1972). *Minerva: An Electronic Town Hall*. Policy Sciences, 3(4), 457-474.
- Gregory, S.** (2010). *Cameras Everywhere: Ubiquitous Video Documentation of Human Rights, New Forms of Video Advocacy, and Considerations of Safety, Security, Dignity and Consent*. Journal of Human Rights Practice, 2(2), 191-207. doi:10.1093/jhuman/huq002
- Hill, C. A., & Helmers, M. H.** (2004). *Defining visual rhetorics*. Mahwah, NJ: Lawrence Erlbaum.
- Huurnink B. et al.** (2010). *Today's and tomorrow's retrieval practice in the audiovisual archive*. Proceedings from CIVR '10: ACM International Conference on Image and Video Retrieval (pp.18-25). Xi'an, China.
- Ito, Mizuko** (2009). *Hanging Out, Messing Around, and Geeking Out*. Cambridge, MA: MIT Press.
- Jamieson, Kathleen** (1988). *Eloquence in an Electronic Age: the Transformation of Political Speechmaking*. New York: Oxford University Press.
- Jarrett, Susan; Losh, Elizabeth & Puente, David** (2006). *Transnational identifications: Biliterate writers in a first-year humanities course*. Journal of Second Language Writing 15, 24-48.
- Jenkins, Henry** (2006). *Convergence culture: Where old and new media collide*. New York, NY: New York University Press.
- Juhasz, Alexandra** (2011). *Learning from YouTube*. Cambridge, MA: MIT Press.
- Kumar, Belhumeur & Nayar** (2008). *FaceTracer: A Search Engine for Large Collections of Images with Faces*. Proceedings from The European Conference on Computer Vision, 340-353.
- Lévy, Maurice & European Commission** (2011). *The New Renaissance*. Brussels: European Commission.
- Losh, Elizabeth** (2008). *Government YouTube: Bureaucracy, surveillance, and legalism in state sanctioned online video channels*. In Geert Lovink & Sabine Niederer (Eds.), *Video vortex reader* (pp. 111-124). Amsterdam, Netherlands: Institute of Network Cultures.
- Losh, Elizabeth** (2009). *Virtualpolitik: An electronic history of government media-making in a time of war, scandal, disaster, miscommunication, and mistakes*. Cambridge, MA: MIT Press.
- Losh, Elizabeth & Alexander, Jonathan** (2010). 'A YouTube of one's Own?': 'Coming out' narratives as rhetorical action. In Christopher Pullen & Margaret Cooper (Eds.), *LGBT identity and online new media* (pp. 23-36). New York, NY: Routledge.
- Losh, Elizabeth** (2011) *Shooting for the public: YouTube, Flickr, and the Mavi Marmara shootings*. In Geert Lovink & Rachel Somers Miles (Eds.), *Video vortex reader II* (pp.

- 283-292). Amsterdam, Netherlands: Institute of Network Cultures.
- Losh, Elizabeth** (2012). *Channeling Obama: YouTube, Flickr, and the social media president*. Comparative American Studies 10(2-3) (forthcoming).
- Loviglio, J.** (2005). *Radio's intimate public: network broadcasting and mass-mediated democracy*. Minneapolis: University of Minnesota Press.
- Lubin, D.** (2003). *Shooting Kennedy: JFK and the culture of images*. Berkeley: University of California Press.
- Lynch, Marc** (2006). *Voices of the new Arab public: Iraq, Al-Jazeera, and Middle East politics today*. New York, NY: Columbia University Press.
- Manovich, Lev** (2001). *The Language of New Media*. Cambridge, Mass.: The MIT Press.
- Manovich, Lev** (2007). *Cultural analytics*. [White paper]. Retrieved from <http://lab.softwarestudies.com/2008/09/cultural-analytics.html>.
- Manovich, Lev** (2009a). *Cultural analytics: Visualizing cultural patterns in the era of 'more media.'* DOMUS (Spring 2009). Milan.
- Manovich, Lev** (2009b). *The practice of everyday (media) life: From mass consumption to mass cultural production?* Critical inquiry 35(2), 319-331.
- Manovich, Lev** (2010). *Software Takes Command*. Italian translation: Milan: Edizioni Olivares, published as Software Culture. Revised version Continuum, 2013.
- Manovich, Lev** (2011a). *From reading to pattern recognition*. In Mieke Gerritzen, Geert Lovink, & Minke Kampman (Eds.), *I read where I am: Exploring the new information cultures*. Amsterdam, Netherlands: Valiz.
- Manovich, Lev** (2011b). *Media Visualization: Visual Techniques for Exploring Large Media Collections*. In Kelly Gates (Ed.), *Media Studies Futures*. Blackwell, forthcoming 2012.
- Manovich, Lev** (2011c). *"What is Visualization?"* Visual Studies, 26(1), 36-49.
- Manovich, Lev and Jeremy Douglass** (2011d). *"Visualizing Change,"* Oliver Grau with Thomas Veigl (Eds.), *Imagery in the 21st Century* (pp. 315-338). Cambridge, MA: The MIT Press.
- Manovich, Lev** (2012). *Trending: The promises and the challenges of big social data*. In Matthew K. Gold (Ed.), *Debates in the digital humanities*. The University of Minnesota Press.
- McCarty, William** (2007). *Beyond retrieval? Computer science and the humanities*. Retrieved from <http://www.mccarty.org.uk/essays/McCarty,%20Beyond%20retrieval.pdf>
- McKinsey Global Institute** (2011). *Big data: The next frontier for innovation, competition, and productivity*. New York, NY: McKinsey Global Institute.
- Navas, Eduardo** (2012). *"Modular Complexity: The Collapse of Time and Space into Search."* AnthroVision journal. France: Lyon, forthcoming Spring/Summer 2012.
- Olson, L. C., Finnegan, C. A., & Hope, D. S.** (2008). *Visual rhetoric: a reader in communication and American culture*. Los Angeles: Sage.
- Oomen, Johan, et al.** (2009). *Images for the future: Unlocking the value of audiovisual heritage*. In J. Trant & D. Bearman (Eds.), *Museums and the Web 2009: Proceedings*. Toronto: Archives & Museum Informatics.
- Panti, Mervi** (2009). *Misfortunes, memories and sunsets: Non-professional images in Dutch news media*. International Journal of Cultural Studies 12(5), 471-489.
- Pasma, Trijntje** (2011). *Metaphor and register variation: The personalization of Dutch news discourse*. Amsterdam: Vrije Universiteit.
- Postman, Neil** (1992). *How to watch TV news*. New York, NY: Penguin Books.
- Rodrigues, Isabel Galhano** (2010). *Gesture space and gesture choreography in European Portuguese and African Portuguese interactions: A pilot study of two cases*. Gesture in embodied communication and human-computer interaction. Berlin: Springer.
- Sarkar, Sudeep, et al.** (2005). *The humanID gait challenge problem: Data sets, performance, and analysis*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(2), 162-177.
- Seinstra, Frank J. et al.** (2007). *High-performance distributed image and video content analysis with parallel-Horus*. IEEE Multimedia 14(4), 64-75.
- Snoek, Cees and Smeulders** (2010). *Visual-concept search solved?* IEEE Computer 43(6), 76-78.
- Snoek, Cees et al.** (2004-2010). *The MediaMill TRECVID 2004-2010 semantic video search engine*. In Proceedings from TRECVID Workshops. Gaithersburg, MD.
- Ubois, Jeff** (2005). *New approaches to television archiving*. First Monday 10(3). Retrieved from <http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/1210/1130>.
- UCLA Film & Television Archive, UCLA MA Program in Moving Image Archive Studies and INA (Institut National de l'Audiovisuel, France) and Ina SUP European Centre for Research, Training and Education on Digital Media** (2010). *Reimaging the Archive: Remapping and Remixing Traditional Models in the Digital Era* [Conference]. Retrieved from <http://polaris.gseis.ucla.edu/reimagining/statement.htm>.
- van de Sande, Koen E. A., Gevers, Theo & Snoek, Cees** (2010). *Evaluating color descriptors for object and scene recognition*. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(9), 1582-1596.
- van de Sande, Koen E. A., Gevers, Theo & Snoek, Cees** (2011). *Empowering visual categorization with the GPU*. IEEE Transactions on Multimedia 13(1), 60-70.
- Yamaoka, S., L. Manovich, J. Douglass, F. Kuester** (2011). *"Cultural Analytics in Large- Scale Visualization Environments."* IEEE Computer, December 2011.
- Zepel, Tara** (2011). *Culture as data as culture : Re-presenting the videostyle of the 2008 presidential campaign ads*. In Geert Lovink & Rachel Somers Miles (Eds.), *Video vortex reader II* (pp. 234-239). Amsterdam, Netherlands: Institute of Network Cultures.

A Sense of Place: Mapping Fictional Landscapes in Literary Narratives

Lynch, John

University of California, Los Angeles, US

Kurtz, Wendy

wendy@itc.humnet.ucla.edu

University of California, Los Angeles, US

Rocchio, Michael

University of California, Los Angeles, US

Maps are one of the most universal forms of communication. Having no need for a written language or alphabet, their ability to convey meaning lies within the images that are drawn on the canvas, be it a cave wall, a piece of parchment, or a computer screen. This paper investigates how digital mapping tools present students and readers of literature with an unprecedented ability to map fictional spaces in their own liking and how these various representations can influence our understanding of and relationship with literary works. Typically thought of as a tool to assist the user in finding a certain location, when used to identify fictional spaces, maps possess the power to convey societal and cultural ideologies. In recent scholarship, there has been a growing interest in the interplay between maps and narrative. "Story maps," "fictional cartography," and "geospatial storytelling" are some of the terms utilized to describe the relationship between place and space (Caquard). Our paper takes the current discussion of the connection between narrative and space to another plane by exploring the exchange between fictional spaces and narrative as represented by maps. We will construct the the following spaces conceived of by Spanish and Latin American novelists through immersive, 3D mapping technology: Mocondo, the setting for many of Gabriel García Marquez's writings loosely based off of his hometown in Columbia. Obaba, a town continually developed in the mind of Basque writer Bernardo Axtaga for his novels and short stories; the village of Ordial in the mythical region of Celama generated by Luis Mateo

Díez as the setting for his novels in the Leonese countryside; Región - the territory invented by Juan Benet for his novelistic trilogy; and Clarín's (Leopoldo Alas y Ureña) *Vetusta* - a reimagining of the Asturian capital, Oviedo - in what many scholars consider the most important Spanish novel of the 19th century, *La Regenta*. By focusing our work on a defined time period and culture it provides us data to not only analyze how fictional space is described and mapped, but also how it can be culturally determined.

Often the inspiration for fictional maps comes from literature and their creation is not necessarily left solely to cartographers. Botticelli's depiction of Dante's Inferno portraying the various stages of hell is one of the more iconic examples of a fictional map of popular literature, depicting the various stages of hell as described by Dante. The image, however, while rooted in the words of Dante, also serves to reinforce the modernity of Renaissance society through its geometry of order and symmetry, a trend in many works of the Renaissance. By depicting the chaotic world of hell as seemingly rational and ordered, it allowed Renaissance society to interpret Dante as a corollary to their contemporary world, or at least their ideal vision of it (Padron). The maps that accompany J.R. Tolkien's Lord of the Rings novels are a much more recent example of this phenomenon in which maps of literature are utilized to bridge the gap between fantasy and real world ideals. Tolkien uses his maps to instill the idea of tranquility being found in the eastern world (the Shire) and evil situated in the western world (Mordor), a direct correlation to current thought in the wake of World War I (Croft). The relationship between the fictional world of literature and a "real space" provides a powerful tool that allows readers of these works to interpret the creator's words in a way which attempts to link them to the real world and prevailing ideologies (Piatti).

This line between the space and place of the real world and that of the imaginary world becomes increasingly blurred in the last half century (Joliveau). The works of Henri Lefebvre and Ed Soja contribute to the dissolution of concrete boundaries between the materiality of the physical space and the abstraction of the mental space. They theorize that the influence of ideas, signs, and texts on spaces is just as influential as the materiality of a space (Lefebvre, Soja). This grants imaginary spaces another sphere of influence upon the real world and begs for further examination into how readers interpret literary spaces and what impact they may have on their understanding of real spaces.

Fortunately, new digital tools in mapping and geospatial analysis allow for a more thorough and in-depth evaluation of these fictional spaces of literature. They provide the opportunity for the individual reader to craft the imaginary space in their own image instead of defaulting to the omniscient representation provided by the cartographer, artist, or scholar. Digital tools are becoming increasingly more intuitive and their interfaces and operability are more user friendly. This contributes to their use by an ever-growing audience and allows for more rapid production with less years of rigorous training. Innovations in digital tools also provide various ways in which spaces can be mapped. GIS provides a solution to the problematic of shifts in places over time. A particular location or space is not static and spaces evolve over time. These mutations could be caused by natural or man-made disasters--such as an earthquake or tsunami in the former and war in the latter--or simply by natural evolution over time. In literature, characters often move between space and place, at times in a non-linear fashion. When attempting to visualize changes to fictional landscapes over time, digital mapping technologies resolve many complexities by allowing the reader to create thick maps that look at time and space through layered pieces.

GIS-based tools allow users to map spaces in more conventional methods, while other applications that integrate technologies such as Google Earth provide users with a platform to map not only spaces, but also plot journeys that characters take through these spaces. Other technologies take this one step further and grant the ability to create more immersive 3D environments that attempt to generate a viewpoint similar to what the literary character actually experiences. Furthermore procedural modeling software such as ESRI CityEngine allow for rapid and adaptable modeling

of large spaces with written rules, a skill that was previously reserved for the domain of those skilled in modeling software, which is labor intensive. Couple this with new software capable of creating large and expansive terrains, such as VUE, and a user has all of the necessary tools to create expansive imaginary spaces unique to their interpretation of a reading. Finally one may take this even further by incorporating these individually designed environments into amateur gaming software, such as Unity, and share them with other users from around the globe to take them on an immersive tour of the space. This new way of developing imaginary spaces provides critical insight into literary works and opens new avenues of study, which were previously unavailable.

In addition to examining how GIS tools enhance the reader's experience with literary texts, we will discuss the importance of combining cartography with literature in the classroom. The combination of digital mapping with literary studies allows us to maintain some of the fundamental pedagogical principles of the humanities, such as close reading and attention to detail. The interpretation of literary texts to create digital maps that represent imaginary or fictional locations, often rooted in real spaces, results in the translation of a text into a new medium. Preparing a digital map of a fictional space requires students to critically examine a narrative in order to recreate a visual representation (via a map) of the space described. Working with GIS platforms shows students how to interact with literary texts in a new way, while teaching them digital mapping techniques and new skills in computer-mediated learning. George Siemens' learning theory entitled "connectivism" helps underline the importance of combining humanistic study with digital technologies. The connectivist's theory incorporates the ways we ingest information and learn in the digital age, where as behaviorism, cognitivism, and constructivism ("three broad learning theories most often utilized in the creation of instructional environments") ignore the technical advances that now greatly impact the way in which students learn. (Siemens).

References

- Caquard, Sébastien.** (2013). *Cartography I: Mapping narrative cartography*. Progress in Human Geography. 37.1 (February 2013): 135-144.
- Heasley, Lynne.** (2013). *Shifting Boundaries on a Wisconsin Landscape: Can GIS Help Historians Tell a Complicated Story?* Human Ecology. 31.2 (June 2003): 183-213.
- Piatti, Barbara, Bär, Hans Rudolf, Reuschel, Anne-Kathrin, Hurni, Lorenz, Cartwright, William,** (2009), *Mapping Literature: Towards a Geography of Fiction*. Cartography and Art. Springer Berlin Heidelberg. 1-16.
- Siemens, George** (2005). *Connectivism: A Learning Theory for the Digital Age*. International Journal of Instructional Technology and Distant Learning. 2.1 (January 2005): np.
- Lefebvre, Henri.** (2005). *The Production of Space*. Oxford: Blackwell.
- Soja, Edward** (1996). *Thirdspace: journeys to Los Angeles and other real-and-imagined spaces*. Oxford: Blackwell.
- Padron, Ricardo.** *Mapping Imaginary Worlds* (2007). Maps: Finding Our Place in the World edited by James Akerman and Robert Karrow. Chicago & London: University of Chicago Press.
- Croft, Janet Brennan.** *The Great War and Tolkien's Memory: An Examination of World War I Themes in the Hobbit and Lord of the Rings*. (2002). Mythlore 23.4: 4-21.
- Joliveau, Thierry.** *Connecting Real and Imaginary Places through Geospatial Technologies: Examples from Set-jetting and Art-Oriented Tourism*. (2009). The Cartographic Journal 46.1 (February 2009): 36-45.

Sentiment Analysis for the Humanities: the Case of Historical Texts

Marchetti, Alessandro

amarchetti@fbk.eu

Fondazione Bruno Kessler

Sprugnoli, Rachele

sprugnoli@fbk.eu

Fondazione Bruno Kessler / University of Trento

Tonelli, Sara

satonelli@fbk.eu

Fondazione Bruno Kessler

1. Introduction

In this paper we investigate the possibility to adapt existing lexical resources and Natural Language Processing (NLP) methodologies related to Sentiment Analysis (SA) to the historical domain.

Sentiment analysis aims at the computational treatment of opinion, sentiment and subjectivity in texts.¹

Current research in SA mainly focuses on the identification of sentiment and opinions in areas such as social media², news^{3 4}, political speeches⁵, customer and movie reviews^{6 7 8}. To our knowledge, SA in the context of the humanities has been rarely explored^{9 10 11}.

Many SA tools often take advantage of polarity lexicons, i.e. a lexicon of positive and negative words and n-grams. In a polarity lexicon, each word is associated with its *prior polarity*, i.e. the polarity of the word out of the context. A SA system uses these lexicons to evaluate the polarity of a whole text, a sentence or a topic within a text. The availability of a sentiment lexicon is thus a crucial step toward the creation and training of any SA application. Unfortunately, the majority of existing SA lexicons are for English (e.g. Harvard General Inquirer¹²) while no lexicon for Italian has been developed yet.

The polarity of a word can however be different according to its context of use. A word can be negated and change its polarity ('ice-cream is good' vs 'ice-cream is *not* good') or have different usages ('they fought a terrific battle' vs 'I loved the film, it was *terrific!*'). To account for these differences, a system must be able to handle the *contextual polarity* of a word, i.e. the different polarity of a word according to its syntactic, semantic or pragmatic context^{13 14 15 16}.

Apart from manual annotation or automatic mapping from English, crowdsourcing methodologies can offer a viable solution to collect a polarity lexicon¹⁷ and to annotate a large dataset¹⁸.

The need to explore the application of SA to historical texts has emerged thanks to the collaboration between the authors and the *Italian-German Historical Institute* (ISIG) in Trento. This collaboration is aimed at developing tools that can help historians access and understand textual data through the adoption of NLP methods. In particular, SA has been identified as notably relevant to quantify the general sentiment of single documents, to track the attitude towards a specific topic or entity over time and across a large collection of texts, and to allow specific search based on sentiment. This is crucial, for instance, to research on the history of ideology, evolution of political thought, etc.

The dataset used for our research is the complete corpus of writings of Alcide De Gasperi, one of the founders of the Italian Republic, made of about 3,000 documents and 3,000,000 words.

Using this corpus as a case study, two experiments have been carried out and are described in this paper. The aim of these experiments is the evaluation of i) how existing lexical resources for SA perform in the historical domain and ii) the feasibility of a sentiment annotation task for historical texts either with expert annotators and crowdsourcing contributors.

2. Prior Polarity Experiment

The first experiment on De Gasperi's corpus has been carried out using two existing polarity lexicons, namely SentiWordNet

¹⁹ and WordNet-Affect²⁰, to calculate the prior polarity of lemmas and measure the general sentiment of each document within the corpus. The goal was to test how resources built on contemporary languages can deal with historical texts.

SentiWordNet and WordNet-Affect have the great advantage of being extensions of a well-known resource called WordNet²¹. This allowed us to map the word senses (called *synsets*) with a positive, negative or neutral polarity in SentiWordNet and WordNet-Affect to the corresponding Italian synsets in MultiWordNet²², in which Italian synsets are aligned with WordNet ones. At the same time, lemmas were automatically extracted from De Gasperi's corpus using the TextPro tool²³: the total of 70,178 lemmas was reduced to 36,304 after excluding lemmas that can't have a polarity score (e.g. numbers, articles). Each lemma was then automatically associated with the most frequent synset in MultiWordNet and its polarity score: this association covered 14,874 lemmas (40.97%) among which 9,650 were neutral. This process, followed by a manual check of the scores, produced a list of 5,224 lemmas with a polarity score: 449 with an absolute positive score (e.g. '*giubilo/rejoicing*'), 576 with an absolute negative score (e.g. '*affranto/broken-hearted*') and the others with an intermediate score.

The general sentiment of each document in the corpus was finally calculated summing up the polarity scores of the lemmas appearing both in the documents and in our list, and visualized through a gauge diagram in the A.L.C.I.D.E. web platform [dh.fbk.eu/projects/alcide-analysis-language-and-content-digital-environment] (Figure 1).



Fig. 1: Document visualization: sentiment and key-concepts

Historians' evaluation of the results was positive for most of the documents but a more specific need emerged: historians are indeed more interested in the polarity of a specific topic and in its evolution over time, rather than in the global polarity of a document that can give us indications only about the general sentiment conveyed in it. However, as historical texts are complex documents in which several topics can be identified, the global polarity of the document is not enough to identify the polarity of a single topic.

To address these requirements, we performed the experiment presented in Section 3 aimed at annotating SA at the level of topic in De Gasperi's corpus, following a contextual polarity approach.

3. Crowdsourcing Experiment for Contextual Polarity

In order to perform a pilot experiment, we identified two topics which were relevant in De Gasperi's writings, namely "sindacato"¹⁹ (*trade union*) and "sindacalismo" (*trade-unionism*).

A corpus of 525 sentences was automatically extracted from De Gasperi's corpus, where each sentence contained at least one of the two lemmas "sindacato" and "sindacalismo". The previous and the following sentence were added as a context as well. Each sentence was annotated by two expert annotators, while a third annotation was collected through the crowdsourcing platform CrowdFlower [www.crowdflower.com] after performing a majority voting over 5 judgements.

The two expert annotators were asked to create gold standard data (GS), i.e. a set of sentences on which both annotators gave the same judgements, from a subset of the corpus (60 sentences, 11% of the whole corpus). Both expert annotators and crowdsourcing contributors were then

asked to annotate the contextual polarity of the two topics in the sentences with one of the four possible judgments (i.e. positive, negative, neutral, unknown) given a simple set of instructions and some annotation example.

In addition to the manual annotation, we also calculated the prior polarity for each sentence using the same algorithm applied to the documents and described in Section 2.

1. The feasibility of this task was then evaluated calculating:

1. 2. the accuracy of the crowdsourced annotation over GS (figure 2), i.e. how well non-expert contributors performed the task;
2. 3. the accuracy of the prior polarity for each sentence over GS (figure 2), i.e. how well the Italian prior polarity lexicon performed on the sentences in comparison to the contextual polarity approach;
3. the inter-annotator agreement (IAA) with the Fleiss's kappa measure (figure 3)²⁴, i.e. the level of consensus between the annotators.

Accuracy	Crowd	Prior
Overall	68.30%	43%
Negative	55.50%	22%
Neutral	80%	31%
Positive	46.60%	86%
Unknown	0	n.a.

Fig. 2: Accuracy scores

IAA	Fleiss' kappa
annotator A vs annotator B vs crowd	0.39
annotator A vs annotator B	0.46
annotator A vs crowd	0.35
annotator B vs crowd	0.35

Fig. 3: IAA results

The overall accuracy score for the crowd-collected judgements in Figure 2 (68.3%) indicates the general complexity of the task. In particular negative and positive polarities are more difficult to identify (55.5% and 46.6%) than neutral polarity (80%).

Considering the prior polarity scores in Figure 2, we observe that accuracy is always lower than in the crowd annotation setting, except for the positive judgements (86%).

The IAA agreement in Figure 3 confirms that SA is a challenging task²⁵. The highest kappa-score is found if we consider the two expert annotators (0.46), but it is not much higher than the situation in which we consider 3 annotators (0.39) or one of the two experts and the crowd judgement (0.35). In general, the type of documents have a great influence on the agreement scores: past works report that news stories can achieve an agreement of 0.81²⁶, whereas social media (tweets) can be as low as 0.321²⁷.

4. Conclusions and Future Works

This paper presented two experiments related to SA and involving a corpus of historical texts. In the first one we created a new Italian lexical resource for sentiment analysis starting from two existing lexicons for English and we applied it to

measure the polarity of an entire document using a prior polarity approach. In the second experiment, the use of crowdsourced annotation to obtain contextual polarity of a specific topic was exploited.

The long term goal of our ongoing research is to create a system to support historical studies, which is able to analyze the sentiment in historical texts and to discover the opinion about a topic and its change over time.

In the near future we plan to perform domain adaptation of existing annotation schemes developed for SA^{28 29} and of the Italian lexical resource we created. Particular attention will be devoted to a step-by-step evaluation by historians in order to tailor the results of our work to their needs.

References

- Pang, B. and Lee, L.** (2008). *Opinion mining and sentiment analysis*, *Foundations and Trends in Information Retrieval* 2 (1-2) , 1-135.
- Basile, V. and Nissim, M.** (2013). *Sentiment analysis on Italian tweets*, Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pp. 100–107, Atlanta, United States.
- Wiebe, J., Wilson, T., and Cardie, C.** (2005). *Annotating Expressions of Opinions and Emotions in Language*, *Language Resources and Evaluation* 39 (2/3) , 164-210.
- Amer-Yahia, S., Anjum, S., Ghenai, A., Siddique, A., Abbar, S., Madden, S., Marcus, A. and El-Haddad, M.** (2012). *MAQSA: a system for social analytics on news.*, in K. Selçuk Candan; Yi Chen 0001; Richard T. Snodgrass; Luis Gravano and Ariel Fuxman, ed., 'SIGMOD Conference', ACM, , pp. 653-656.
- Somasundaran, S. and Wiebe, J.** (2010). *Recognizing Stances in Ideological On-line Debates*, in Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, pp. 116-124, Los Angeles, CA. Association for Computational Linguistics.
- 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, pp. 168-177 .
- Pang, B. and Lee, L.** (2004). *A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts*, in Proceedings of the 42nd annual meeting on Association for Computational Linguistics, Barcelona, ES , pp. 271-278 .
- Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C.** (2011). *Learning word vectors for sentiment analysis*, in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, pp. 142-150.
- Cooper, D. and Gregory, I. N.** (2011). *Mapping the English Lake District: a literary GIS*. Transactions of the Institute of British Geographers, 36: 89–108.
- Kakkonen, T. and Kakkonen, G.G.** (2011). *SentiProfiler: Creating Comparable Visual Profiles of Sentimental Content in Texts*, in Proceedings of the Workshop on Language Technologies for Digital Humanities and Cultural Heritage, pp. 62–69, Hissar, Bulgaria. www.aclweb.org/anthology/W11-4110.
- Mohammad, S.** (2011). *From Once Upon a Time to Happily Ever After: Tracking Emotions in Novels and Fairy Tales*, in Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, pages 105–114, Portland, OR, USA. Association for Computational Linguistics. www.aclweb.org/anthology/W11-1514.
- Stone, P.** (1997). *Thematic text analysis: new agendas for analyzing text content*, in Carl Roberts, ed., *Text Analysis for the Social Sciences*, Lawrence Erlbaum Associates, Mahwah, NJ .
- Kim, S. M., and Hovy, E.** (2004). *Determining the sentiment of opinions*, in Proceedings of the 20th international conference on Computational Linguistics.
- Wilson, T., Wiebe, J., and Hoffmann, P.** (2005). *Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis*, in Proceedings of the conference on human language

- technology and empirical methods in natural language processing, pp. 347-354.
- Nasukawa, T. and Yi, J.** (2003). *Sentiment Analysis: Capturing Favorability Using Natural Language Processing*, in Proceedings of the Conference on Knowledge Capture (K-CAP).
- Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., and Potts, C.** (2013). *Recursive deep models for semantic compositionality over a sentiment treebank*, in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1631-1642.
- Mohammad, S. and Turney, P. D.** (2013). *Crowdsourcing a Word-Emotion Association Lexicon*. Computational Intelligence 29 (3) , 436-465.
- Pang, B. and Lee, L.** (2004). *A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts*, in Proceedings of the 42nd annual meeting on Association for Computational Linguistics, Barcelona, ES , pp. 271-278 .
- Baccianella, A. E. S. and Sebastiani, F.** (2010). *SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining*, in Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10).
- Strapparava, C. and Valitutti, A.** (2004). *WordNet-Affect: An affective extension of WordNet*, in Proceedings of the 4th International Conference on Language Resources and Evaluation, pp. 1083-1086.
- Fellbaum, C.**, ed. (1998). *Wordnet, an Electronic Lexical Database*, MIT Press.
- Pianta, E., Bentivogli, L. and Girardi, C.** (2002). *MultiWordNet: developing an aligned multilingual database*, in Proceedings of the First International Conference on Global WordNet.
- Pianta, E., Girardi, C. and Zanoli, R.** (2008). *The TextPro Tool Suite*, in Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06).
- Artstein, R., and Poesio, M.** (2008). *Inter-coder agreement for computational linguistics*. Computational Linguistics, 34(4), 555-596.
- Pang, B. and Lee, L.** (2008). *Opinion mining and sentiment analysis*, *Foundations and Trends in Information Retrieval* 2 (1-2) , 1-135.
- Balahur, A., and Steinberger, R.** (2009). *Rethinking Sentiment Analysis in the News: from Theory to Practice and back*. Proceeding of WOMSA.
- Basile, V. and Nissim, M.** (2013). *Sentiment analysis on Italian tweets*. Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pp. 100-107, Atlanta, United States.
- Wiebe, J., Wilson, T. and Cardie, C.** (2005). *Annotating Expressions of Opinions and Emotions in Language*, Language Resources and Evaluation 39 (2/3) , 164-210.
- Di Bari, M., Sharoff, S., and Thomas, M.** (2013). *SentiML: functional annotation for multilingual sentiment analysis*. In Proceedings of the 1st International Workshop on Collaborative Annotations in Shared Environment: metadata, vocabularies and techniques in the Digital Humanities.

Circling around texts and language: towards 'pragmatic modelling' in Digital Humanities

Marras, Cristina
 cristina.marras@cnr.it
 Lessico Intellettuale Europeo e Storia delle Idee

Ciula, Arianna
 ariannaciula@gmail.com
 University of Roehampton, London, UK

1. Introduction (1)

Digital Humanities (hereafter DH) is surely a field in rapid evolution, where open questions (2) are numerous and self-reflexivity is not new. This paper aims to contribute to the discussion around general and somehow obvious questions - debated globally, often daily, by scholars - by reformulating the issues at stake in the following terms: other scholarship in the humanities and techno-sciences (McCarty 2013b¹) has passed/ passes via the experimental and the formal; in what way is the experimental and the formal done in DH different and similar? To address this question the argument will circle around two interrelated concepts:

- that of 'text', intended in its processuality (Meister 2007²) and in a wide sense from linear to discontinuous narrative, from manuscripts to printed editions, encompassing hybrid modalities such as maps (Eide 2013a³) and 'narrative drawings' (Groensteen 2012⁴);
- that of 'modelling', intended as DH-specific research and teaching activity (modelling rather than models; see McCarty 200⁵, Jannidis and Flanders 2013⁶), but also connected to multifaceted conceptualisations of modelling (in particular Kraleemann and Lattmann 2013⁷) as used and seen by other disciplines and practices.

Firstly, ways in which new technologies and languages influence approaches to texts and the consequences for research will be discussed by recalling the results of some research projects (e.g. AGORA⁸, Marras and Lamarra 2013⁹, Marras 2013b¹⁰) and by referring to the literature that reflected at large on the influence the creation and use of, for example, standard markup languages have on the relation between scholars and texts (e.g. Buzzetti 2002¹¹).

This has an important epistemological consequence that in the paper will connect directly the reflection on one of, or possibly the privileged object/s of humanities research - text – to the second focus of the argument: by modelling knowledge we somehow provide for an abstract way of looking at the world. Is DH research prone to privilege a symbolic analysis of texts in opposition to a pragmatic one?

Despite being informed by modelling as rooted in computing and therefore in mathematical reasoning, modelling in DH is directly interconnected to the work on texts, but is not only a way to see patterns of similarity across texts (e.g. Eide 2013b¹²). Beyond what is branded as DH research, other imaginative practices of assembling 'toolkits' to find patterns in human production exist (e.g. Hockney 2006¹³). Can we sustain the unicity of the lens of computing, of modelling through computing? While being a rather theoretical exercise, a comparative perspective on modelling would need to be experimental in its nature: is modelling in computing going to lift our way of seeing and therefore thinking to another level of analysis?

While keeping these questions in mind, we note that significant things are happening. DH or humanities computing in its former vest (Schreibman et al. 2004¹⁴) is being institutionalized and is an example of how cross-border fertilisation, namely interdisciplinarity, is possible (e.g. McCarty 2013a¹⁵, Marras 2012¹⁶, Ciula 2013¹⁷). We acknowledge that DH scholars and professionals are making use of a 'blended' style (intended as in McCarty 2009¹⁸) reflected both in their language, by adopting expressions to account for a new scenario, and in their research and teaching approaches, often integrating computational methods and terminologies with modes of discourse associated to more established scholarship in the humanities (Marras 2013a¹⁹, Flanders 2009²⁰).

DH opened the stage to doing and talking about research recurring to innovative and diverse knowledge as well as to an innovative and diverse way of organising and conceptualising it. Software has also been developed to explore and represent current networked knowledge configurations (e.g. Lima's 'Knowledge Atlas', Quagliotto's 'Knowledge Cartography'²¹). However, a (cultural) change to inform the sharing of practices and results is possibly still lacking behind (Ciula 2013²²). Specific modelling practices in DH could be lead to

combine *theoria cum praxis*, anchoring innovative approaches to a solid theoretical framework.

2. Pragmatic Modelling

In this paper we claim that language as mediation in designing and contextualising models is crucial. We do so by focusing on the concept of pragmatic modeling (4) intended as research strategy, framed within the complex cognitive, social, and cultural functioning of DH practices affected by cross-linguistic and interdisciplinary dimensions. Pragmatic modelling is understood as being anchored to theory and language, while at the same time claiming some freedom from both (e.g. in digital textual editing one might adopt the OHCO model while at the same time questioning it deeply). It operates within the relational and dynamic aspects of modeling.(5)

Middle out method and metaphoric reasoning/language

Metaphors as meta-models (Kralemann and Lattmann 2013²³) and linguistic tools are called to explore the creative power of pragmatic modelling (Getachew 2006²⁴ and Mazzocchi-Fedeli 2013²⁵) and to move forward the theoretical reflection as framed so far. With this respect, metaphorical reasoning exemplifies a specific strategy to guide modelling in DH: pragmatic modelling can be understood as facilitating a middle-out approach based on metaphorical language. In general terms, this translates into a move away from the dichotomy at interplay between bottom-up (models emerging from particulars or 'artifacts of study') and top-down (models imposed on particulars) approaches to what we propose to call a middle-out method; a method that acts at the crossing point of data and models adapting itself to specific "textual contexts". Three interrelated properties (adapted from Verschueren, 1995) can be associated with this method:

- Variability - the range of choices in the use of language cannot be seen as static in any respect.
- Negotiability - such choices are not made mechanically or according to strict rules or fixed form-function relationships, but on the basis of highly flexible principles and strategies, thus also implying the indeterminacy and unexclusiveness of the choices made.
- Adaptability - such negotiable choices can be adapted based on specific needs and contexts according to the variable range of possibilities.

DH Practices

We will narrow down our approach with selected case studies that show how in DH (in comparison to other contexts) choices are not made mechanically or according to fixed theories, but on the basis of flexible principles and strategies potentially open to creative reasoning. Indeed, with respect to modelling theorised in computer sciences, the challenge in DH is to shift the lens of computing up the scale, to embrace the experimental nature of modelling at the lower level of the scale (e.g. in computing coding) and see indeed how it can scale up (e.g. to do critical scholarship with/via it).

The opportunities enabled by modelling (e.g. emergence of patterns of relation, behaviour, and shape) are rooted not only in a 'demonstrative' and 'literal language' but also in the metaphorical one (Marras 2013²⁷, McCarty 2006²⁸). Some modelling attempts in DH and cultural heritage formalisations more in general have embarked in dedicated efforts to problematise terminology (e.g. CIDOC-CRM²⁹; Pundit³⁰); some prominent DH scholarship is reflecting on the limits of adopting uncritically the language of computer sciences (e.g. Eide et al. 2013³¹ on spatio-temporal concepts in humanities and arts; Simpson et al. 2013³² on what a 'person' is in ontological models for the humanities; Renear 2013³³ on what 'datasets' are for libraries, publishing, data curation, and DH); some other DH scholarship has ventured in creative attempts at establishing neologisms (e.g. "factoid" as described in Short

and Bradley 2005³⁴): is there a trade-off in projecting historical lexicons in new contexts of use?

3. Conclusions

In conclusion, departing from open and interdisciplinary conceptualisations of objects of analysis – such as texts – and of certain explorative and epistemological strategies of analysis – such as modelling – the authors will show how the spectrum of research in DH is indeed expanding our boundaries of knowledge. However, modelling practices are more than often constrained by the language they are embedded in, either because terminology is not problematised enough or because language is not used imaginatively.

Notes

(1) This paper is the result of an intense discussion carried out between the two authors in the last years on the nature of DH research practices and strategies. This discussion took place in different contexts, in particular during the work on DH infrastructures carried out at ESF (Moulin et al. 2011³⁵), but also as part of informal exchanges stemming from diverse teaching experiences and works within collaborative research projects in the broad area of DH. Recently a discussion focused on modelling within the forum of Humanist (see references³⁶) triggered some further reflections partially formalised in this paper.

(2) "Is there such thing as DH? Is DH unique in its practice and research strategy?" See Gold 2012³⁷ for a rich overview on this discussion.

(4) We subscribe to a functional perspective on the study of language. By focusing on use, a pragmatic perspective is also integrative in that it aims at encompassing the full complexity of the cognitive, social, and cultural functioning of language (Verschueren 1995³⁸).

(5) We mean both the interplay between the object of analysis and the model (usually referred as mapping; e.g. Kralemann and Lattmann 2013, 3417³⁹), as well as across different levels of the interpretative process (e.g. close and distant reading, symbolic/sintagmatic and semantic/paradigmatic levels of text analysis).

References

1. McCarty, W. (2013b). DH Conference 2013, Busa Award lecture 2013, Alliance of Digital Humanities Organizations: www.youtube.com/watch?v=nTHa1rDR680 (accessed 3 March 2014)
2. Meister, J. Ch. (2007). *Events are Us, Amsterdam International Electronic Journal for Cultural Narratology*, (AJCN) 4/Autumn 2007: cf.hum.uva.nl/narratology/a07_meister.htm (accessed 3 March 2014)
3. Eide, Ø. (2013a). *Why Maps Are Silent When Texts Can Speak. Detecting Media Differences through Conceptual Modelling*. Buchroithner M. F. et al. (eds), Proceedings of the 26th International Cartographic Conference, Dresden 2013: www.icc2013.org/_contxt/_medien/_upload/_proceeding/31_proceeding.pdf (accessed 3 March 2014)
4. Groensteen, T. (2012). *Définitions. L'Art de la bande dessinée*. Citadelles & Mazenod: 17-75.
5. McCarty, W. (2005). *Humanities computing*. Basingstoke: Palgrave Macmillan.
6. Jannidis, F. and Flanders, J. (2013). *A concept of data modelling for the humanities*. DH Conference 2013: dh2013.uni.edu/abstracts/author.html?q=author%3A22jannidis%2C20fotis%7C%7Cjannidis%2C%20Fotis%22 (accessed 3 March 2014)
7. Kralemann, B. and Lattmann C. (2013). *Models as icons: modeling models in the semiotic framework of Peirce's theory of signs*, Synthese, November 2013, Volume 190, Issue 16: 3397-3420.

- 8. AGORA, Scholarly Open Access Research in European Philosophy**, CIP-Pilot Action (CIP-ICT-PSP-2010-4, n. 270904: www.project-agora.org
- 9. Marras, C. and Lamarra, A.** (2013). *Scholarly Open Access Research in Philosophy: Limits and Horizons of a European Innovative Project*. DH Conference 2013: dh2013.uni.edu/abstracts/ab-316.html (accessed 3 March 2014)
- 10. Marras, C.** (2013b). *AGORA-Scholarly Open Access Research in Philosophy*. An overview'. Beiträge zur Geschichte der Sprachwissenschaft, Projektberichte, 23.2: 295-301.
- 11. Buzzetti, D.** (2002). *Digital Representation and the Text Model*, New Literary History, 33, 1, Winter 2002: 61-88 | 10.1353/nlh.2002.0003
- 12. Eide, Ø.** (2013b). Ontologies, data modelling, and TEI. TEI Conference 2013: digitlab2.let.uniroma1.it/teiconf2013/program/papers/abstracts-paper#C107 (accessed 3 March 2014)
- 13. Hockney, D.** (2006). *Secret knowledge: rediscovering the lost techniques of the old masters*. London: Thames & Hudson.
- 14. Schreibman, S.; Siemens, R.; Unsworth, J.** (eds.) (2004). *Companion to Digital Humanities*. Oxford: Blackwell, www.digitalhumanities.org/companion (accessed 3 March 2014)
- 15. McCarty, W.** (2009). *Being reborn: the humanities, computing and styles of scientific reasoning*. New Technology in Medieval and Renaissance Studies 1: 1-23. Invited paper originally delivered at the Renaissance Society of America conference, Clare College, Cambridge, 7 April 2005.
- 16. Marras, C.** (2012). *Pragmatica e comunicazione: orizzonti interdisciplinari*. Gensini, S. and Forgione L. (eds.) Filosofia della comunicazione. Roma: Carocci.
- 17. Ciula, A.** (2013). *Which Changes are Currently Taking Place in our Research and Academic Culture?*, International colloquium: Research Conditions and Digital Humanities: What are the Prospects for the Next Generation?, German Historical Institute in Paris, June 2013: www.slideshare.net/arimare/presentation-ciulaparis2013 (accessed 3 March 2014)
- 18. McCarty, W.** (2009). *Being reborn: the humanities, computing and styles of scientific reasoning*. New Technology in Medieval and Renaissance Studies 1: 1-23. Invited paper originally delivered at the Renaissance Society of America conference, Clare College, Cambridge, 7 April 2005.
- 19. Marras, C.** (2013a). *Structuring multidisciplinary knowledge: Aquatic and terrestrial metaphors*. Knowledge Organization, Mazzocchi, F. and Fedeli, G. (eds.), special issue, Fall 2013: 392-399.
- 20. Flanders, J.** (2009). *The Productive Unease of 21st-century Digital Scholarship*, Digital Humanities Quarterly, vol.3, 3, 2009: www.digitalhumanities.org/dhq/vol/3/3/000055/000055.html (accessed 3 March 2014)
- 21. Knowledge cartography:** www.knowledgecartography.org (accessed 3 March 2014)
- 22. Ciula, A.** (2013). *Which Changes are Currently Taking Place in our Research and Academic Culture?*, International colloquium: Research Conditions and Digital Humanities: What are the Prospects for the Next Generation?, German Historical Institute in Paris, June 2013: www.slideshare.net/arimare/presentation-ciulaparis2013 (accessed 3 March 2014)
- 23. Marras, C.** (2013b). *AGORA-Scholarly Open Access Research in Philosophy*. An overview'. Beiträge zur Geschichte der Sprachwissenschaft, Projektberichte, 23.2: 295-301.
- 24. Getachew, M-T.** (2006). *Metaphor*. Continuum Encyclopedia of British Philosophy. Goulder, N., Grayling A.C., Pyle, A. (eds.). London: Thoemmes Continuum.
- 25. Mazzocchi, F. and Fedeli, G.** (eds.) (2013). *Special Issue: 'Paradigms of Knowledge and its Organization: The Tree, the Net and Beyond'*, Knowledge Organization, vol. 40, n.6.
- 26. Verschueren, J.** (1995, revised 2012). *The pragmatic perspective. Handbook of Pragmatics*, Östman J-O. & Verschueren J. (eds.), John Benjamins: Amsterdam/Philadelphia.
- 27. Marras, C.** (2013a). *Structuring multidisciplinary knowledge: Aquatic and terrestrial metaphors*. Knowledge Organization, Mazzocchi, F. and Fedeli, G. (eds.), special issue, Fall 2013: 392-399
- 28. McCarty, W.** (2006). *Tree, Turf, Centre, Archipelago or Wild Acre? Metaphors and Stories for Humanities Computing*. Literary and Linguist Computing, April 2006, 21, 1: 1-13.
- 29. CIDOC-CRM, Conceptual Reference Model:** www.cidoc-crm.org (accessed 3 March 2014)
- 30. Pundit:** www.thepund.it (accessed 3 March 2014)
- 31. Eide, Ø; Grossner, K.; Berman, Merrick L.; Ore, C.-E.** (2013). *Issues in Spatio -Temporal Technologies for the Humanities and Arts*. Panel at DH Conference 2013: dh2013.uni.edu/abstracts/ab-319.html (accessed 3 March 2014)
- 32. Simpson, J.E.; Brown, S.; Goddard, L.** (2013). *A Humanist Perspective on Building Ontologies in Theory and Practice*, DH Conference 2013: dh2013.uni.edu/abstracts/ab-413.html (accessed 3 March 2014)
- 33. Reaner, A.** (2013). *Text encoding, ontologies, and the future*. Keynote at TEI Conference 2013: digitlab2.let.uniroma1.it/teiconf2013/program/keynotes/ (accessed 3 March 2014)
- 34. Short, H. and John B.** (2005). *Texts into Databases: The Evolving Field of New-style Prosopography, Literary and Linguistic Computing*, 2005, 19.5: 3-24
- 35. Moulin, C.; Nyhan, J.; Ciula, A; et al.** (2011). *Research Infrastructures for the Digital Humanities*, ESF Science Policy Briefing 42, Sept. 2011: www.esf.org/publications (accessed 3 March 2014)
- 36. Humanist Discussion Group**, vol. 27, n. 471: 27.471 models of computation: www.digitalhumanities.org/humanist (accessed 3 March 2014)
- 37. Gold, M.K., ed.** (2012). *Debates in Digital Humanities*. University of Minnesota Press: dhdebates.gc.cuny.edu/ (accessed 3 March 2014)
- 38. Verschueren, J.** (1995, revised 2012). *The pragmatic perspective. Handbook of Pragmatics*, Östman J-O. & Verschueren J. (eds.), John Benjamins: Amsterdam/Philadelphia.
- 39. Kraleemann, B. and Lattmann C.** (2013). *Models as icons: modeling models in the semiotic framework of Peirce's theory of signs*, Synthese, November 2013, Volume 190, Issue 16: 3397-3420.

STAK – Serendipitous Tool for Augmenting Knowledge: Bridging Gaps between Digital and Physical Resources

Martin, Kim

kimberleymartin@gmail.com
University of Western Ontario

Greenspan, Brian

brian_greenspan@carleton.ca
Carleton University

Quan-Haase, Anabel

aquan@uwo.ca
University of Western Ontario

The book as a research tool is not obsolete, but it is being augmented by surrounding clouds of data. Web research lacks access to analog and archival sources, while those who access physical materials often do so in isolation from other information resources. Despite the relevance of the physical library for humanities research, its once uncontested position has been challenged by the vast, easily accessible resources available on the Internet. Kemman et al. (2013) found that even among humanities scholars in the Netherlands and Belgium, general search systems are predominant (e.g., Google and JSTOR)¹. Within these systems, searching with simple keywords is the primary search strategy employed by users, while more advanced search strategies are uncommon. The physical library, then, no longer exists in a vacuum; rather, the browsing experience is constantly being augmented by scholars' perceptions and expectations of the virtual information that supplements the information contained in the physical copies of books. Library users accustomed to working in digital environments are eager to have additional information

such as keywords, tables of contents, links to other literature, and thematic suggestions at their fingertips. Yet, humanities scholars rarely critically examine the intricate relation between the physical and digital protocols of information storage, access, and organization.

To examine the relation between the physical library and online resources, we conducted a series of user studies in which we employed a head-mounted compact action video camera to record the experiences of participants as they navigate the physical space of the library². Two important findings emerge from this work. First, participants report that physical stacks are better suited for facilitating the discovery of new resources than the catalog. As one participant reported, "I feel that if I'm looking up a topic and I go to the shelf for videogame studies, I'll find books that I hadn't seen in the catalog, but in the general field; even if I lose track of a single book, I'll find another in the general area; this happens to me constantly." Second, our user studies have shown that going to the physical library to search and browse for books is still an important and rewarding component of students' research.

The careful organization of information in libraries and archives, together with the tangible, physical presence of analogue documents, has given humanist scholars a level of comfort in their primary research space. These scholars allude time and again to the chance encounter with information as a welcome digression within the library stacks, so much so that this phenomenon has become a normal, even anticipated aspect of their research process^{3 4 5}. While serendipity is an elusive concept, a number of models have recently been proposed that decompose the phenomenon into specific facets. Not all models of serendipity emphasize the same facets; however, the majority of models tend to stress the relevance of "noticing" or discovering novel information as a central component to serendipity. This component is also central to the work of humanists as they navigate stacks, archives, and even digital environments.

One of the first models of serendipity, proposed by Erdelez (2004), underscored the shift in attention that occurs when a person is solving a problem or engaged in a foreground activity and then suddenly encounters unrelated information⁶. In this model of information encountering, noticing novel information is central to any kind of engagement with information. Similarly, in their model of serendipity in everyday chance encounters, Rubin et al. (2011) describe how the act of noticing is central to any discovery of new resources⁷. Likewise, Makri and Blandford's (2012) study demonstrates that new mental connections or insights result from encountering novel, unexpected information⁸. This finding has been sustained by our own initial user studies, in which participants consistently reported finding additional, unexpected information when searching for a specific book in the stacks.

Humanist scholars in general note the importance of serendipity, with Hoeflich (2007) comparing them to "ancient mariners" who "set out on voyages of discovery hopeful of finding new lands of milk and honey" (p. 813). For Hoeflich, Walpole's two-part definition of serendipity as a union of accident and ability was missing a third part: opportunity. Hoeflich notes the importance of the tactile aspects of archival documents, arguing that significant qualities of historical artifacts can be lost when reproduced in a different medium. The loss of serendipity is often perceived as resulting from the move to digital environments, and away from deep engagement and interactions with physical books, manuscripts, and archival records, as Martin and Quan-Haase (2013) have shown⁹. They found that serendipity was central to the work of historians, because the encounter with a single key resource on library shelves or in archives could significantly affect the outcome of their research. The historians interviewed by Martin and Quan-Haase noted that they tend to limit their use of digital environments to quick searches, browsing subject headings, and fact-checking until they can recreate experiences of chance encounters with material that is similar to that experienced in library stacks or archival collections.

Bess Sadler, Manager, Software Engineering Team at Stanford Library, spoke about what she believes is missing from the process of library discovery in a blog post adapted

from her keynote for ACCESS 2012. Sadler (2012) writes about what keeps scholars motivated to work, how they consider library browsing a pleasurable part of their research process, and that she believes there is "something missing" from the virtual library collection. She states, "I think we're still falling short of a full replacement for physical browsing. I think we're still falling short of providing the kind of emotional, physical, and spatial sensory experience that shelf browsing, at its best, can provide"¹⁰. Sadler's call for change is pertinent to any scholar who has benefitted from research in a library setting. She came to this conclusion after conducting a study with graduate students in the humanities and social sciences, in which she noticed their love of browsing physical stacks, and their inability to conduct their research in the same way in a virtual environment. Even since Sadler's talk, there have been several attempts by digital humanists to remediate the chance encounter or serendipitous experience (serendip-o-matic.com, mechanicalcurator.tumblr.com)

We will describe our own attempts to reconcile the physical and digital library spaces through the creation of a dedicated mobile app, STAK (Serendipitous Tool for Augmenting Knowledge). Built upon Greenspan's web-based geolocative StoryTrek authorware^{11 12}, this web app will be grounded in both the theoretical models of serendipity research and in user tests conducted with digital humanities scholars. STAK will use a combination of RFID and wifi triangulation to detect a researcher's location in the library stacks with fine-grained precision. After determining the user's proximity to specific subject headings, it will automatically retrieve full-text excerpts on the same subject from networked databases, along with reviews, relevant titles by nearby authors, and different works drawn from both open-source and proprietary databases, sorting these hits on-the-fly by type and relevance. The app will link to live datasets that refresh dynamically, putting vast repositories of research networks into the user's hand, wherever she may be. STAK will bring the shelves to life by providing ready access to the growing clouds of virtual data that surround physical texts, effectively annotating the library's physical holdings of books, journals, manuscripts, prints, blueprints, microfiche, films and videos.

We will describe the results of our intermediate tests, in which we ask users to perform library search tasks using a simulation of the proposed STAK browser app while wearing a head-mounted action camera and speaking aloud (though softly) about their user experiences. Our test app (Figure 1) is designed to simulate the function of STAK without relying on actual geolocation by retrieving online information from a constrained dataset relevant to the user's pre-determined location in the stacks. Each test is followed up with post-test interviews in which we ask users to elaborate on their experiences of STAK's perceived functionality, accuracy, and user interface design.

We will detail how these tests are allowing us to develop a user model and search algorithms based on the following three questions:

1. Why is the library the perfect setting for serendipitous discovery? What makes this setting so conducive to the chance encounter?
2. How do the serendipitous experiences of library users reflect the findings of previous models of this phenomenon?
3. How does the physical, controlled environment affect the possibility of chance encounters of information?

The attempt by digital humanists to replicate serendipity cannot rely on online sources alone. We need to bridge existing notions of humanities research and its inherited practices with current information resources and digital tools. When fully implemented, STAK will allow users to combine their physical and digital navigational tactics in a single search environment. Books, no longer the end point of the research process, become the fiducials that allow access to a wider cloud of information.



Fig. 1: Figure 1 - Simulated STAK Browser App

References

1. Kemman, M., Kleppe, M., & Scagliola, S. (2013). *Just Google it - digital research practices of humanities scholars*. Cornell digital libraries. doi: arxiv.org/abs/1309.2434
2. Quan-Haase, A., Greenspan, B., & Martin, K. (2013). *Look Around: Linking Bits to Books*. In HASTAC. Toronto. Retrieved from hastac2013.org/schedule-2/brian-greenspan/
3. Duff, W. M., & Johnson, C. A. (2002). *Accidentally Found on Purpose: Information-Seeking Behavior of Historians in Archives*. The Library Quarterly, 72(4), 472–496. doi:10.2307/40039793
4. Hoeflich, M. H. (2007). *Serendipity in the stacks, fortuity in the archives* *. Law Library Journal, 99(4), 813–827.
5. McClellan III, J. E. (2005). *Accident, luck, and serendipity in historical research*. Proceedings Of The American Philosophical Society, 149(1), 1 – 21. Retrieved from www.jstor.org/stable/4598905
6. Erdelez, S. (2004). *Investigation of information encountering in the controlled research environment*. Information Processing & Management, 40(6), 1013–1025. doi:10.1016/j.ipm.2004.02.002
7. Rubin, V. L., Burkell, J., & Quan-Haase, A. (2011). *Facets of serendipity in everyday chance encounters: a grounded theory approach to blog analysis*. Information Research, 16(3). Retrieved from informationr.net/ir/16-3/paper488.html
8. Makri, S., & Blandford, A. (2012). *Coming across information serendipitously – Part 1: A process model*. Journal of Documentation, 68(5), 684–705. Retrieved from www.emeraldinsight.com/journals.htm?articleid=17051067
9. Martin, K., & Quan-Haase, A. (2013). *Are e-books substituting print books? Tradition, serendipity, and opportunity in the adoption and use of e-books for historical research and teaching*. Journal of the American Society for Information Science and Technology, 64(5), 1016-1028.
10. Sadler, E. (Bess). (2012). *Brain Injuries, Science Fiction, and Library Discovery*. Solvitur ambulando. Retrieved February 10, 2013, from www.ibiblio.org/bess/?p=248
11. Greenspan, B. (2011). *Songlines in the Streets: Story Mapping with Itinerant Hypertext*. In R. Page & B. Thomas (Eds.), New Narratives: Theory and Practice (pp. 153–169). Lincoln, Nebraska: University of Nebraska Press.
12. Khaled, R., Barr, P., Greenspan, B., Biddle, R., & Vist, E. (2011). *StoryTrek: Experiencing Stories in the Real World*. In Proceedings of Mindtrek. Tampere, Finland.

Designing the next big thing: Randomness versus serendipity in DH tools

Martin, Kim

kimberleymartin@gmail.com

University of Western Ontario

Quan-Haase, Anabel

aquan@uwo.ca

University of Western Ontario

A number of recent initiatives within the DH community promote the design, development, and implementation of digital tools aimed at speeding up, clarifying, or otherwise improving the research practices of humanities scholars. This year, the One Week | One Tool (OWOT) summer institute, funded by the National Endowment for the Humanities, resulted in the creation of Serendip-o-matic, a serendipity engine for digital research. This tool relies on users to feed it a selection of text or citations in order to create a list of keywords, which it then uses to find related information. The documents returned are taken from the Digital Public Library of America (DPLA), Europeana, and Flickr¹. The participants of the 2013 OWOT initiative are not alone in their quest to design a digital tool geared toward enhancing the chance encounter with information, resources, ideas, research materials, and even people. Tim Sherratt, the manager of Trove at the National Library of Australia, often includes an element of chance in the tools he designs for use in the humanities. For instance, in his tool Trove News Bot, Sheratt (2013) allows users to interact with a Twitter stream by sending tweets with directions (such as #luckydip), which will return random results from the National Archives of Australia's digital collection². Similar tools have been developed that introduce serendipity into the collections of the DPLA and the British Library.

One motivation for the development of digital tools aimed at enhancing serendipity in digital environments comes out of the need to redesign and recreate the complexity of the research environment found in library stacks and archival collections. It is often argued that this complexity may be lost in digital environments, which are highly predictable and primarily based on keyword search. To what extent serendipity is reduced in digital search is debatable. Nonetheless, this perception of loss directly affects how scholars, and in particular humanities scholars, adopt and use digital tools. A study of historians' research practices suggests that these scholars are skeptical of conducting their research exclusively in digital environments because they lack the ability to encounter key resources (primary and secondary materials) that could have a major impact on their research findings³. In this study, the authors also found that historians were willing to experiment with digital tools, if these could recreate opportunities for encountering information. Hence, scholars perceive the discovery of resources, browsing, and chance encountering as central elements of their research practice that can, and need to, be supported online.

Outside of academia, a number of tools have emerged that try to introduce serendipity into the online experience. What is less clear from the literature is how to best support this process, as a wide range of approaches have been suggested ranging from interactions in social media⁴, exploration in non-search related digital environments, and information search in digital environments⁵. The approach most commonly taken is to introduce serendipity into the online information search experience; this is often done by introducing some element of randomness and thereby reducing the predictability of search results. An example of this approach is BananaSlug, which returns random results to a search query. Other approaches include reversing or modifying the ranking in which search results are presented online⁶. This would draw attention to a different set of items because users commonly tend to investigate only the first and perhaps second pages of search results. All of these approaches aim at broadening "the search space, promoting encounters with items that might not, under

existing algorithms, come to the attention of the user". While the majority of digital tools aimed at promoting serendipity have emerged outside of the humanities, a series of tools have recently been developed with humanists in mind. These tools have garnered considerable attention in the field, but it remains unclear what element of serendipity they support. Part of the problem is the fact that the concept of serendipity is elusive⁷ and difficult to pinpoint. Reducing it to the introduction of randomness, however, does not seem to be the best way to move forward, even though it is the one most commonly utilized. A second problem, and perhaps more concerning, is that scholars need to first understand that serendipity is not a one-dimensional concept but, rather, includes a number of related facets, which need to inform tool design and implementation. The present paper critically examines four DH tools that encourage serendipitous results and attempts to place these within current models of serendipity:

Serendip-o-matic (<http://serendip-o-matic.com/>)
Trove News Bot (<https://github.com/wragge/trovenewsbots>)
Mechanical Curator (<http://mechanicalcurator.tumblr.com/>)
DP.LA Bot (<https://github.com/wragge/trovenewsbots>).

As a basis for this examination, we have established the main facets of serendipity obtained from the extensive literature in Library and Information Science (LIS). Through this comparative study, we aim to accomplish two goals. First, there is a gap in understanding exactly what aspects of serendipity digital tools support. By merging the literature in LIS with tool design in DH, we hope to create greater clarity as to what aspects have been supported. Second, the results of the study will determine what future developments are needed to better support the work of humanists in digital environments.

Interviews with 20 history scholars inform the first phase of this study. These scholars indicated a desire for serendipitous encounters with material to remain a part of their research process after the integration of digital texts to their work. After discovering that historians were seeking new methods of information acquisition online, further interviews were conducted with DH scholars to see what methods they were using to browse information. The results of these two sets of qualitative data will be discussed and used to demonstrate a need for a serendipity tool within the DH community.

The second phase of this research is an in-depth exploration of the four information-discovery tools listed above. These tools will be examined in terms of Erdelez's (2004) model of information encountering outlined below⁸. After analyzing each tool carefully, follow-up interviews will be conducted with the creators of each tool to discuss their intentions for and reflections upon, the use of the tool by humanist scholars.

A wide range of models of serendipity have been developed relying on very different data sets and assumptions. Erdelez (2004) developed one of the first models and emphasized the experience of information encountering (IE), which she defined as a type of opportunistic acquisition of information. Erdelez's (2004) utilized an experimental setting, where participants were asked to look for information related to a foreground problem and the researcher observed how they would react to information related to a background problem. As part of her model, Erdelez (2000) identified five elements:

noticing: the perception of encountered information;
stopping: the interruption of the initial information seeking activity;
examining: the assessment of usefulness of the encountered information;
capturing: the extraction and saving of the encountered information for future use; and
returning: the reconnection with the initial information seeking task.

In Erdelez's (2004) model, a person is primarily focusing on the information needs related to a foreground problem. However, cues related to another problem, a background problem, may catch the person's attention. If the person notices the cues and stops to examine the newly encountered information, then there is an opportunity for discovering unexpected resources. It is this process of noticing, examining, and capturing that digital tools try to emulate or support.

Each of the four tools reflects one or more aspects of the serendipitous process as outlined by Erdelez, (see Table 1).

	Noticing	Stopping	Examining	Capturing	Returning
Serendip-o-matic		✓	✓		
Trove News Bot	✓	✓	✓	✓	
Mechanical Curator	✓	✓	✓	some	
DP.LA Bot	✓	✓	some		

The tools listed above, with the exclusion of Serendip-o-matic, select materials randomly and then present these to followers on Twitter. Randomness, as we know, does not necessarily mean that serendipity will occur. These tools all provide links to places that users can go to receive extraneous materials in the hopes that something of interest will come their way.

Interestingly, the capturing element of these tools seems to be largely disregarded. Considering the DH community is acutely aware of the need to instantly capture digital documents and the associated metadata with citation tools (Zotero), none of the examined tools includes this element in their framework. This leads the authors to conclude that future design could focus on this element of capturing information, and could introduce a method that allows for the saving of documents so users can retrace their footsteps after returning to the initial task or foreground problem. Our critical analysis of various DH tools and how they support serendipity provides opportunity to further enhance these tools as well as a means to design additional tools that can impact the research practices of humanities scholars.

References

1. CHMN. (2013). *Serendip-o-matic: Let your sources surprise you*. One Week | One Tool. Retrieved October 31, 2013, from serendip-o-matic.com/about
2. Sherratt, T. (2013). *Conversations with Collections*. discontents. Retrieved October 31, 2013, from discontents.com.au
3. Martin, K., & Quan-Haase, A. (2013). *Are e-books substituting print books? Tradition, serendipity, and opportunity in the adoption and use of e-books for historical research and teaching*. Journal of the American Society for Information Science and Technology, 64(5), 1016-1028.
4. Bogers, T., & Björneborn, L. (2013). *Micro-serendipity: Meaningful coincidences in everyday life shared on Twitter*. In Proceedings of iConference (pp. 196–208).
5. Quan-Haase, A., Burkell, J., & Rubin, V. L. (n.d.). *The role of serendipity in digital environments*. In Encyclopedia of Information Science and Technology. IGI Global.
6. Jansen, B. J., Spink, A., & Saracevic, T. (2000). *Real life, real users, and real needs: A study and analysis of user queries and on the web*. Information Processing & Management, 36(2), 207–227.
7. Merton, R. K. (2004). *The travels and adventures of serendipity: a study in sociological semantics and the sociology of science*. Princeton, N.J.: Princeton University Press.
8. Erdelez, S. (2004). *Investigation of information encountering in the controlled research environment*. Information Processing & Management, 40(6), 1013–1025. doi:10.1016/j.ipm.2004.02.002
9. Erdelez, S. (2000). *Towards understanding information encountering on the Web*. In Proceedings of the 63rd annual meeting of the American Society for Information Science (pp. 363–371). Medford, N.J.: Information Today.

Small-Scale Big Data: Experimental Literature and Distributed Computing

Mauro, Aaron

mauro@uvic.ca

University of Victoria, Electronic Textual Cultures Lab

Introduction

This short paper describes how a small-scale implementation of big data text analysis can be used for reading single texts and testing algorithmic processes. By using a small Hadoop cluster, distributed computing methods can be used to parse typographically experimental texts and delineate even minute units of meaning. This experiment takes Ron Silliman's 26-volume collection entitled *The Alphabet* (1979–2008) as an object of study because it anticipates and resists quantitative analysis techniques. Silliman describes his brand of "language writing" as a kind of "composition as investigation" that requires active participation of the reader to make meaning.¹ The process of running resistant texts through an algorithmic system also exposes the interpretive biases of computational methods of reading poetry. Large scale text analysis can reveal promising texts for further inquiry, but these systems can also be used on a small-scale to complement qualitative human readings with quantitative results. Matthew Jockers has recently warned in *Macroanalysis* (2013) that "from thirty thousand feet, something important will inevitably be missed. The two scales of analysis, therefore, should and need to coexist."² This paper shows how a proximate use of "distant reading" can render a text productively unfamiliar and also show how distributed computing systems can be used to structure highly experimental literary texts.

Background

The core of my method is animated by two observations: First, Google has become a constant touchstone for DH scholarship. The proceedings for the 2013 meeting of DH in Lincoln, Nebraska includes 156 references to Google Search or other Google products, including Anna Jobin and Frederic Kaplan's attempt to reverse engineer Google's autocomplete algorithms.³ However, the scale and sophistication of Google's systems make them impenetrable for all but the most experienced Google employees responsible for building these proprietary systems. Second, there has been a recent boom in avant-garde or experimental writing that is fuelled by the growing awareness of large scale text analysis. I argue, therefore, that experimental literature has begun to function as a resistant dataset that can test and even react to algorithmic methods. This new experimental corpus thereby engages in a reciprocal critique of both literary and technical systems. My experiment is grounded in contemporary cultural and technological contexts, while seeking to explore the relationship between corporate analysis tools and literary production. The technologies that support this "Big Data Revolution" were developed by many of the most important technology companies in the US, with Google, Yahoo, and Facebook among them.⁴ Hadoop was derived from Google's MapReduce and Google File System white papers. Doug Cutting (Yahoo!) and Mike Cafarella (Google) spun off an open source implementation in 2005 that is now released by Apache.⁵⁶ Hive was written by Facebook to streamline the process of writing MapReduce jobs in Java by allowing queries to be written in SQL. A secondary purpose for this research explores the capabilities and weaknesses of proprietary systems from their open source derivatives.

Implementation

A small-scale implementation of Hadoop and Hive on Amazon Web Services Elastic Map Reduce (EMR) platform is a highly reliable and cost effective means of working with distributed computing methods. As both a pedagogical and rapid experimental tool, EMR automates the networking between data nodes throughout the system and opens the scalability of Hadoop to an extremely broad user base. Hive allows for queries to be written in HiveQL, which does not follow the full SQL-92 standard but rather retains many of the features of MySQL.⁷ While the distributed nodes are often virtualized on EMR and the user has no way to determine the full composition of the cluster, the costs remain extremely low and do not require complex discussions with system administrators in a restrictive institutional environment.⁸ Hadoop's speed and flexibility is the result of its very simple order of operations and physical architecture that allows for scaling from just three to potentially thousands of nodes.⁹ As the corpus size expands, this system can scale rapidly with only modest additional investment of research funds.

Analysis

The term "distant reading" first articulated by Franco Moretti has come to describe the process of quantitative analysis.¹⁰ It arose out of a very human inability to read all the texts that compose World Literature. A remarkable number of issues emerge from this very simple state of affairs. Firstly, we now have a kind of writing that acts like reading. In other words, the SQL queries used to parse and structure vast strings of data represents a readable trace of reading. Coding commands for this system represents a "composition as investigation."¹¹ Secondly, the scalability of these systems allows for machine reading systems to perform human lifetimes worth of reading tasks in mere minutes or hours. The computer's ability to perform repetitive tasks at speeds also means that the interpretive experience of texts is now outside the domain of human perception. "The Hadoop Distributed File System" white paper described the pace of operations through the "heartbeats" that guide the operations of the distributed machines by explaining how the TaskTracker "can process thousands of heartbeats per second."¹² Thirdly, this technological moment represents the collapse of the false dichotomy between philosophical, subjective, and speculative analysis, and scientific, objective, empirical analysis. Quantitative methods have the potential to erase the long held doctrine of the "two cultures" that divides the sciences and humanities and presumes that "intellectuals, in particular literary intellectuals, are natural Luddites."¹³

Conclusion

The use of the word "experimental" carries a strategic significance in this context. It is a word that simultaneously accesses literary, scientific, and technological discourses. My experiment is primarily animated by the "aesthetic provocation" of avant-guard writing.¹⁴ Because computation relies upon symbolically stable inputs, making computational sense of non-sense characters becomes a central challenge to overcome in the study of contemporary experimental literature with computational methods; in order to read these non-sense characters, the "stop list" for this topic model may need to comprise the entire corpus of proper words. Rather than treating literature as strictly quantifiable data that algorithmic analysis can simply glean information from, I propose a methodology that assumes that literary information is profoundly resistant, reactive, and unpredictable. Kenneth Goldsmith claims, in *Uncreative Writing* (2011), that "digital media has set the stage for a literary revolution."¹⁵ While Goldsmith is thinking here about distribution methods on the Web, there is little doubt that literature is responding to the technological context into which it is published. It is now time to include the algorithm in this history of the avant-guard.

Acknowledgements

I would also like to thank the generous support of the Electronic Textual Cultures Lab at the University of Victoria and Implementing New Knowledge Environments group. This work is supported by the Social Science and Humanities Research Council of Canada.

References

1. **Silliman, Ron** (1984). *For L=A=N=G=U=A=G=E*. The Language Book. Eds. Bruce Andrews and Charles Bernstein. Southern Illinois UP. 14. Print.
2. **Jockers, Matthew** (2013). *Macroanalysis: Digital Methods and Literary History*. Urbana: U of Illinois P. 9. Print.
3. **Broeckmann, Andreas**. *Digital Culture, Art, and Technology*. IEEE Multimedia 12.4 (Oct.-Dec. 2005): 9-11. Web. 8 April 2013.
4. **Mayer-Schönberger and Kenneth Cukier** (2013). *Big Data: A Revolution that Will Transform how We Think Live, Work, and Think*. New York: Houghton Mifflin Harcourt. Print.
5. **Shvachko, Konstantin, Hairong Kuang, Sanjay Radia, Robert Chansler** (2010). *The Hadoop Distributed File System*. Symposium on Massive Storage Systems and Technologies and Co-located Events. IEEE: Computer Society. Web. 22nd Oct. 2013.
6. **Dean, Jeffrey and Sanjay Ghemawat**. (2004). *MapReduce: Simplified Data Processing on Large Clusters*. USENIX Association OSDI '04: 6th Symposium on Operating Systems Design and Implementation. Web. 22nd Oct. 2013.
7. Apache Hive. hive.apache.org/
8. Amazon Elastic MapReduce (Amazon EMR). aws.amazon.com/elasticmapreduce/
9. Apache Hadoop. hadoop.apache.org/
10. **Moretti, Franco**. *Conjectures on World Literature*. New Left Review 1 (Jan./Feb. 2000): 54-68. Web. 30 July 2013.
11. **Silliman, Ron**. (2008) *The Alphabet*. Tuscaloosa: U Alabama P. 1057. Print.
12. **Shvachko, Konstantin, Hairong Kuang, Sanjay Radia, Robert Chansler**. *The Hadoop Distributed File System*. Symposium on Massive Storage Systems and Technologies and Co-located Events. IEEE: Computer Society, 2010. Web. 22nd Oct. 2013.
13. **Snow, C.P.** (1961) *The Two Cultures and the Scientific Revolution*. New York: Cambridge UP. 23. Print.
14. **Drucker, Johanna** (2009). *SpecLab: Digital Aesthetics and Projects in Speculative Computing*. Chicago: U of Chicago P. xi. Print.
15. **Goldsmith, Kenneth** (2011). *Uncreative Writing*. New York: Columbia UP. 15. Print.

Pushing Back the Boundary of Interpretation: Concept, Practice and Relevance of a Digital Heuristic

Meister, Jan Christoph
jan-c-meister@uni-hamburg.de
 University of Hamburg

Jacke, Janina
janina.jacke@uni-hamburg.de
 University of Hamburg

heureCLÉA (www.heureclea.de) is a BMBF-funded eHumanities project¹ which combines the two conceptual perspectives on object annotation that set apart the humanities and the 'hard sciences': strictly rule-based explication of uncontroversial object features as exemplified in the measurement of values, such as length, height, density, etc., versus the hermeneutic response for which observation and emotive engagement from a subjective point of view must go hand in hand in order to facilitate interpretation.

These are of course ideal types and the actual practice of annotation in the humanities is situated at their interface, which is heuristics—the methodologically controlled 'art of finding' that goes beyond pure measurement, but whose purpose it is to generate relevant questions rather than conclusive answers. Against this backdrop heureCLÉA aims to implement a digital heuristics module for the text annotation tool CATMA (www.catma.de) so that we may benefit from synergies between the computationally automated and the subjectively motivated, human generated annotation of texts.

1. The heureCLÉA Project

The practical backbone of heureCLÉA are two software developments: *HeidelTime*, which was developed at Heidelberg University, is a rule-based system for the extraction and normalization of temporal expressions.² It needs to be significantly modified to cope with the complexity of literary narratives.³ – The *CATMA* (*Computer Aided Textual Markup & Analysis*) markup tool was developed at Hamburg University. The current release of CATMA (version 4.0) is open source and provides a robust web based annotation environment for collaborative markup.⁴ CATMA not only supports intuitive text annotation in a flexible, XML/TEI-compliant format, but also integrates markup with analytical and visualization functions (cf. Fig.1). This enables users to switch ad hoc between text annotation and text analysis in either direction as well as recursively. CATMA thus supports what Burnard referred to as the 'continuous turning of the hermeneutic wheel'⁵—i.e. the back and forth between formal text analysis and the generation of interpretative hypotheses. In its most recent development phase (called CLÉA)⁶, CATMA then progressed from a stand-alone desktop application to a web-based solution. This adds yet another conceptual dimension, that of collaborative markup.⁷ In CATMA researchers can now share, reuse, amend and dispute each other's markup.

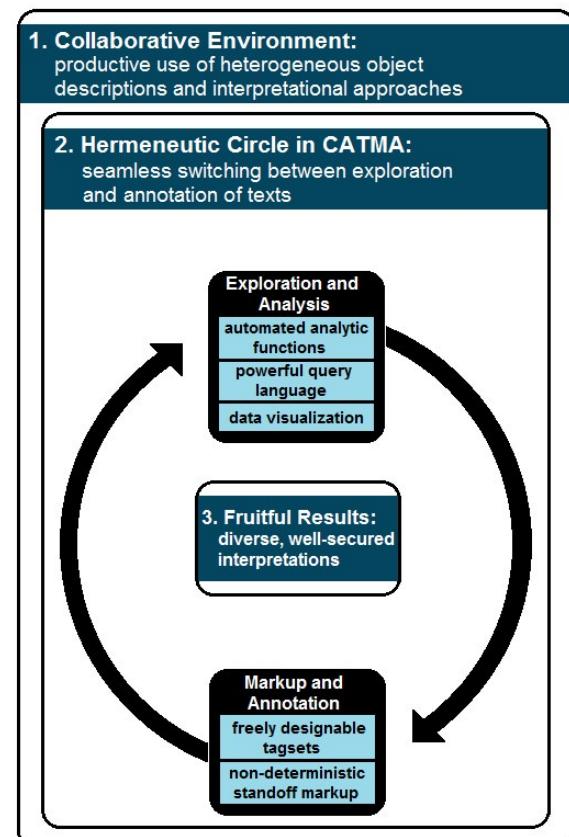


Fig. 1: Exploration and Annotation in CATMA

CATMA's overall design is based on the premise that a DH tool should emulate the methodological and social practice of traditional philology as closely as possible. This practice integrates three methodological primitives: analytical/declarative operations, hermeneutic operations, and discursive critique of explications and theories.⁸ This conceptual high-level design goal also determined our choice between the two competing paradigms of embedded (in-line) markup and external standoff markup, which we regard as methodological rather than technological opposites. Accordingly, embedded markup represent the idea of objective taxonomic universality and the potential immanence of 'truth' in an object. External standoff markup, on the other hand, is based on the acknowledgment of the contingent nature and historical transience of object interpretation. In a contemporary philological perspective, embedded markup therefore constitutes something of a methodological anachronism: for conceptually it resembles the pre-enlightenment model of canonical text exegesis which the modern humanities have long replaced by a critical, self-reflexive hermeneutic approach. Yet this critique is of course of a purely philosophical nature when dealing with pre-interpretive analytical and declarative tasks, such as POS tagging.

However, once higher-level semantics are at stake, these considerations force us to adopt a truly 'hermeneutic' approach to markup.⁹ Interpretation varies depending on interpreter, context and interpretive theory. Accordingly, even elementary markup produced in order to support higher-level interpretation must still remain transparent as one possible account among many, and users must be able to produce and store ambiguous and indeed even contradictory markup for the same text in a standoff manner. Since rich interpretations are best generated in a discursive practice, it is also necessary to enable the easy sharing and combining of markup generated by different interpreters.

However, while these desiderata can all be considered emulative goals which informed the development of CATMA, their conceptual benefit for the digital humanities at large lies elsewhere. A truly non-deterministic and discursive approach to markup yields diverse annotation data—and that type of data can subsequently be analyzed in order to "push back" the boundary separating interpretation and declaration.

What this means is best illustrated by outlining the three components and phases of heureCLÉA (cf. Fig. 2):

1. Narratological Analysis of Temporal Phenomena: In heureCLÉA the identification of the temporal structure of narrative texts is approached concurrently through

- (a) manual collaborative annotation with CATMA, and
- (b) automated temporal tagging with HeidelTime.

In (a) we draw upon (but do not restrict our taggers to) the narratological taxonomy of Genette¹⁰, which is supplemented by a taxonomy suited to capture action and event segmentation. In (b) automated temporal annotations are generated via an UIMA pipeline¹¹ that includes HeidelTime as a rule-based temporal tagger, the TreeTagger¹² as a POS tagger, and Morphisto¹³ for a morphological analysis.

2. Machine Learning Approach towards an Automation of Complex Time Annotation: The next step is the learning of new rules for automated annotation from the manually generated markup. Different methods for the derivation of rules—especially those for typical co-occurrence of temporal expressions—are used. Once integrated into the components of the heureCLÉA UIMA pipeline these rules enable the system to handle more complex annotations. This process is dynamic as growing quantities of markup facilitate more complex modeling strategies based on e.g. distributional approaches (such as Latent Semantic Analysis). Finally, patterns representing typical temporal sequences may be extracted (Sequence Mining).

3. Integration of the Heuristic heureCLÉA Module into CATMA: Once a functional threshold has been passed (i.e.: reliability of automated detection of temporal references of low complexity; performance/robustness) the heureCLÉA UIMA pipeline will be integrated into CATMA as a service. It then provides a 'digital heuristics' for the partially automated, partially interactive generation of temporal markup, and will be tested and evaluated to verify the adequateness of the automatically generated markup (partially through stochastic methods.)

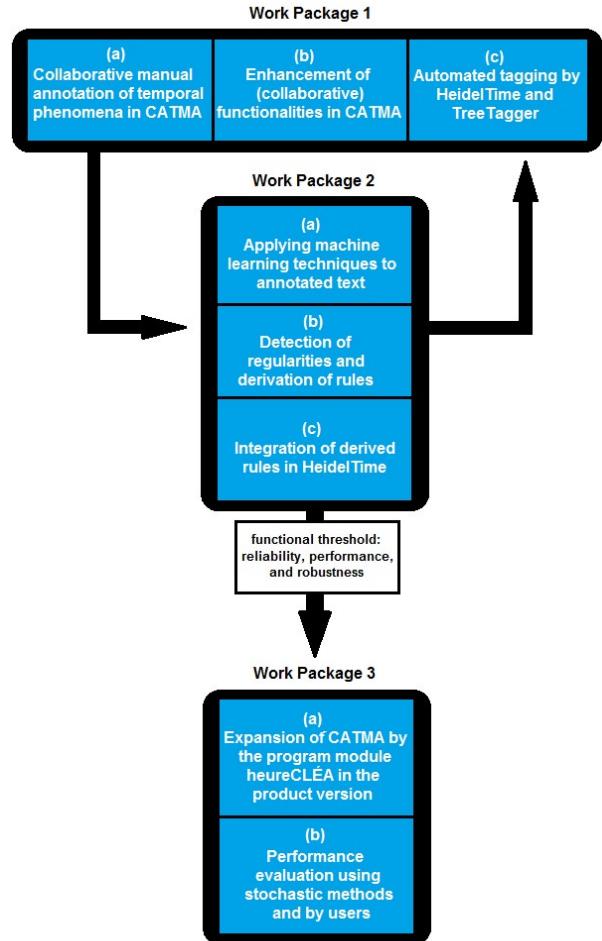


Fig. 2: heureCLÉA's interrelated areas of work

2. The Heuristics/Hermeneutics Divide As a Conceptual Boundary in the Digital Humanities

What is the methodological relevance of this work toward a digital heuristics? The realization that text markup is essentially interpretive *per se* is anything but new.¹⁴ Indeed, the argument about what markup *is* seems somewhat artificial; it might have sufficed to ask literary scholars what markup *is there for*: in their view the *raison d'être* of any object annotation and classification is always interpretation. But is the boundary between a declarative and an interpretive method rigidly defined for the digital humanities?

Some experiences gained in the heureCLÉA project may offer an answer. To begin with, the hermeneutic nature of building time constructs from narratives proves to be only partially owed to the 'fuzziness' of natural language. As a complex symbolic system narrative is also characterized by the intricate coupling of a referential and an indexical semiotics. 'Time' illustrates this: on the surface it is referenced as chronological structure of a particular 'story world'.¹⁵ Yet on a deeper level it is also communicated as an implicit processing instruction encoded in the form of temporal deixis and ellipsis, and of linguistic markers for the contraction or expansion of time (viz. a summary vs. a scenic description). This information is termed 'indexical' because it refers back to the instance of utterance (the narrator or narrating character). The reader needs to process that information in parallel with the referential in order to reconstruct two intersecting chronologies—that of the 'story world', and that of the representational discourse which he is trying to reverse engineer. However, the reach of the indexical extends beyond the text-reader-system: as we read we also become subject to the indexical temporality of the 'how' of representation (*discours*) on an *existential* level.¹⁶ As Ricoeur argues, the reconstruction of temporality on the

referential and discursive levels of narrative is indeed how we learn to experience temporality.¹⁷

Time is only one of many phenomena of narrative representation and understanding characterized by this triple-layered semiotics—and against this backdrop it becomes clear why hermeneutic text interpretation cannot be automated. The machine is (as yet) not an interpreting agent able to engage reflexively and speculatively with its object. It is confined to a fully explicated, operational definition of 'relevance' in terms of known tasks and objectives. While it can resolve an indexical reference in order to compute, say, chronological order and extension, the interpretation of the existential relevance of that double-encoded message as one which also addresses the interpreting mind will remain beyond its ability for as long as it is not equipped with a concept of 'mind'.

However, the foundational human interpretive operations taking place on the elementary declarative and inferential levels **can** in part be approximated statistically through recursive routines. This recursive approximation is the functional characteristic as well as the outer boundary of what we term a 'digital heuristic'. By this we mean a computational tool able to support the heuristic operations of analytical identification and categorization of phenomenological primitives that necessarily precede hermeneutic synthesis. These operations form the indispensable basis of any higher-order 'interpretation'. It is in this intersecting terrain of the hermeneutic and the heuristic that the digital humanities might help to "push back the border" and question the hegemony of interpretation.

References

1. See dbs.ifi.uni-heidelberg.de/index.php?id=129&L=1.
2. **Mazur, P. and Dale, R.** (2010). *WikiWars: A New Corpus for Research on Temporal Expressions*. "Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010)". Massachusetts, pp. 913-922.
3. **Ricoeur, P.** (1983ff). *Time and Narrative (Temps et Récit)*, 3 vols. trans. Kathleen McLaughlin and David Pellauer. Chicago: University of Chicago Press, 1984, 1985, 1988 (1983, 1984, 1985).
4. **Burnard, L.** (2001). *On the Hermeneutic Implications of Text Encoding*. In Fiornonte, D. and Usher, J. (eds.), "New Media and the Humanities: Research and Applications". Oxford: Humanities Computing Unit, pp. 31-38. Available online in the 1998 version at users.ox.ac.uk/~lou/wip/herman.htm [accessed 20 September 2013].
5. "CLÉA" is short for "Collaborative Literature Éxploration & Annotation". The accent deugis was added to highlight the diacritical concerns of Non-Anglo literary scholars. The CLÉA development phase of CATMA was generously supported by two Google Digital Humanities Awards (2010, 2011). For further details see www.catma.de/clea.
6. **Meister, J. C.** (2012). Crowd Sourcing 'True Meaning'. A Collaborative Markup Approach to Textual Interpretation. In **McCarty, W. and Deegan, M.** (eds.), *Collaborative Research in the Digital Humanities. Festschrift for Harold Short*. Ashgate Publishers: Farnham, Surrey/Burlington, pp. 105-122.
7. **McCarty, W.** (1996). *Implicit Patterns in Ovid's Metamorphoses*. "Centre for Computing in the Humanities (CHWP 1996)". projects.chass.utoronto.ca/chwp/mccarty/mcc_8.html (accessed 11 September 2013). First published 1991.
8. **Piez, W.** (2010). *Towards Hermeneutic Markup: an Architectural Outline*. "Digital Humanities 2010. Conference Abstracts". London: Office for Humanities Communication, Centre for Computing in the Humanities, King's College London, pp. 202-205.
9. **Genette, G.** (1972). *Discours du récit*. In id., "Figures III". Paris: Editions Du Seuil, pp. 67-282.
10. **Strötgen, J. and Gertz, M.** (2010). *HeidelTime: High Quality Rule-based Extraction and Normalization of Temporal Expressions*. "Proceedings of the 5th International Workshop on Semantic Evaluation (ACL 2010)". Uppsala, pp. 321-324.
11. **Schmidt, H.** (1994). *Probabilistic Part-of-Speech Tagging Using Decision Trees*. "Proceedings of International Conference on New Methods in Language Processing". Manchester.
12. **Zielinski, A., Simon, C., and Wittl, T.** (2009). *Morphisto: Service-Oriented Open Source Morphology for German*. "State of the Art in Computational Morphology: Workshop on Systems and Frameworks for Computational Morphology (SFCM 2009)". Zürich, pp. 64-75.
13. See among others **Coombs, J. H., DeRose, S. J., and Renear, A. H.** (1987). *Markup systems and the future of scholarly text processing*. "Communications of the ACM", 30 (11): 933-47; **Buzzetti, D.** (2002). *Digital Representation and the Text Model*. New Literary History, 33 (1): 61-88; **Burnard** (2001); **Piez** (2010).
14. **Herman, D.** (2002). *Story Logic: Problems and Possibilities of Narrative*. Lincoln: University of Nebraska Press.
15. **Genette (1972)**.
16. **Ricoeur, P.** (1983ff). *Time and Narrative (Temps et Récit)*, 3 vols. trans. Kathleen McLaughlin and David Pellauer. Chicago: University of Chicago Press, 1984, 1985, 1988 (1983, 1984, 1985).

Visualizing Computational, Transversal Narratives from the World Trade Towers

Miller, Ben

miller@gsu.edu

Departments of English and Communication, Georgia State University

Shrestha, Ayush

ashrestha2@cs.gsu.edu

Department of Computer Science, Georgia State University

Olive, Jennifer

jolive1@gsu.edu

Department of English, Georgia State University

I. Introduction

Following the attacks of September 11, 2001, interviews were conducted with first responders to the World Trade Towers. Each of those 503 interviews describes one witness account of the event – skyscrapers collapsing, choking dust, chaotic communications, fatal desperation. Although focused on individual narratives, each interview corresponds to a larger sequence of events: two massive violations of the right to life that took place in New York City. Given this larger frame, a documentation source dense relative to the event's spatiotemporality, and the evocative, idiosyncratic nature of testimony, we developed a method to computationally elicit and visualize the narratives running across the corpus. This paper describes the work's narratological theory, the visualization developed to facilitate the identification and exploration of transversal narratives, and our analysis of the World Trade Center (WTC) Task Force Interviews.

II. Fabula and witness testimony

As testimony, for ¹, is centered on the body, and the body is contingent with the witness, our analyses focused on the fabular, raw events of the narrative as a foundation for crossdocument coreference. Narrative testimonies like these are, to ², speech acts in the public sphere that serve to solidify a collective memory. Witnessing, consequently, is historiography dependent upon comprehensible dramatic unity as put forward in ³. Narratology indicates that the four primary elements of fabula are events, actors, time, and location ⁴. This corresponds to the events model proposed by ⁵ for human rights violations reporting. Accordingly, we extracted information from the WTC corpus corresponding to these elements, enabling the decontextualization of information from existing narration, the

identification of conflicting, factually questionable accounts, and the visualization of transversal narratives.

III. Applied narratology

The semi-automated information extraction pipeline described in necessitated a method for spatio-temporal interpolation based on keyframing. Events appearing in many narratives, such as the collision of Flight 175 into Tower 2, become key moments for emploting fabula. Given a sufficient number of these events, and of geocodable locations, absolute referents for intermediary material can be interpolated with an accuracy sufficient for correlation. Lookups to gazetteers provided absolute keyframe data.

For extracted Storygrams (computed fabula) of person-place-time without absolute reference, a dramaturgical sequence numbering system was used to emplot all events in the order in which they occurred. These sequence numbers were used to provide interpolated time for events between keytimes. Quotations were used for the named and unnamed entities. Gazetteers were developed to provide absolute values for key times and locations; the global timing list contained 11 events, of which 8 were used. The location list comprised 2,151 names indicating 1,399 unique locations. Many locations lacked a geocodable referent. In those cases, we interpolated the data to suggest a likely location. Recognition of Storygram elements enabled cross-document coreference of implicit entities. Correlation of personal and global data, when visualized, revealed narratives running transversely throughout the corpus.

IV. Narrative visualization

Many tools can visualize temporal aspects of events, notably timeviz.net⁷, and Google charts. Spatial visualization is dominated by mapping tools like Google Maps and Earth, Open Street Maps, ArcGIS, and MapBox. Tools like Google Earth visually animate changes over time. Animations are cognitively demanding, requiring viewers to track changes frame-by-frame. Irregular intervals make this task more problematic.

Multiple interactive 2D synchronous views for time and location are an alternative to animations; operations like zooming, panning, and filtering in one view automatically updates the remaining views. Color and shape are other ways to represent event information. Jern et al. use color coding to link temporal data with spatial data⁸. Color meaning is culturally dependent, so it is not reliably intuitive. In addition, color palettes have to be constrained to allow for visual disambiguation, thereby forcing the system to artificially bin events to a number of discrete color values. Event type diversity should dictate design, not the number of recognizable color options. Shapes present their own challenges⁹. Events can also be shown using a 3D space like Kachina Cube¹⁰. The cube base on the X-Y plane contains a map and the Z-axis represents time. Though 'details-on-demand' techniques can be applied to the visualization, this approach suffers from generic 3D problems like glyph cluttering and scalability. Also, it is often hard in 3D UIs to compare data points in two dimensions (X-Y and Z).

V. Storygraph and narrative

To visualize the narratives and address the issues above, we developed a 2D integrated spatiotemporal visualization called Storygraph¹¹. Storygraph, an extension of parallel coordinates¹², has two parallel vertical axes and an orthogonal horizontal axis. Our novel application adapts this information-rich visualization technique for the presentation of explicit and implicit narrative. The vertical axes represent latitude and longitude and the orthogonal axis represents time. A map location, such as a city or street corner, is represented as a line segment linking the parallel axes. Events occurring at a location are represented by a point on the location line as shown in Figure 1. This technique shows, in 2D, the scope of a corpus,

the relative frequency of documentation at all locations in the corpus, and patterns like co-occurrence in time, co-occurrence at location, or co-occurrence in time and location – one of the unique properties of Storygraph.

Storygraph also facilitates Storylines: linear connections emphasizing the movement of people through the spatiotemporal context of the corpus. Storylines are polyline segments chronologically connecting entities at location. In our implementation, we use dotted lines for storylines to mark the uncertain space between observations.

VI. Visualizations of transversal narratives of 9/11

We applied our visualization to the WTC corpus comprising 17,000 question and answer pairs aimed to elicit first-person narratives of the event. To feed the data into the Storygraph, named and unnamed entities were extracted. This semi-automated process involved much manual verification due to the ambiguity in the natural language.

Gray lines in Storygraph in Figure 1 indicate locations. Each red point shows one Storygram. Black vertical bars indicate global events: the collapse of Tower 1 and Tower 2, the period when people were seen leaping from the towers. The horizontal funnel layout of the points, with the mouth to the left axis, indicates that documentation shows people converging from a wider geographic area to the narrow area around Ground Zero. In essence, Figure 1 shows first responders converging on the scene of two terrorist attacks.

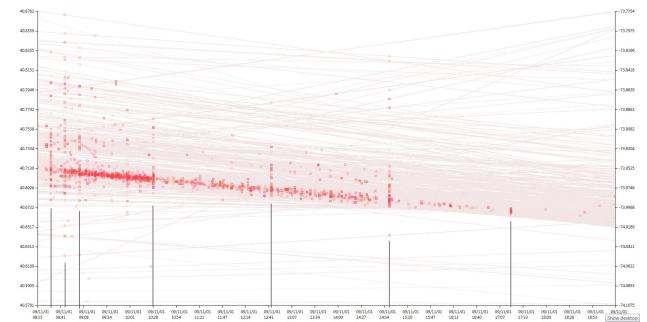


Fig. 1: Storygraph showing all the events extracted from the narratives of the firefighters.

Figure 2 uses the same data to show Storylines of four emergency personnel as they move throughout the spatio-temporal corpus domain. Four features in Figure 2 are of particular importance. First, the geographic domain is highly constrained and covers an area from Staten Island to Central Park. Second, with just 10-20 extracted fabula, a sense of the path of these individuals through the event emerges. The chaotic jumble of points in the period from 8:39 AM to 9:30 AM corresponds to the event's most chaotic moments. Third, the lines of Firefighters Loutsky and Smith stabilize at two locations as they move from emergent crisis to emergency care. And finally, there is the blue storyline of Chief Ganci, which ends at 10 : 12 at 40.71, -74.01, approximately 16 minutes prior to the collapse of WTC Tower 1. Chief Ganci, the highest ranking uniformed fire officer in FDNY, died in that collapse. What Storygraph enables is the identification and organization of fragments from others' statements to reveal the story of what happened to Ganci that day.

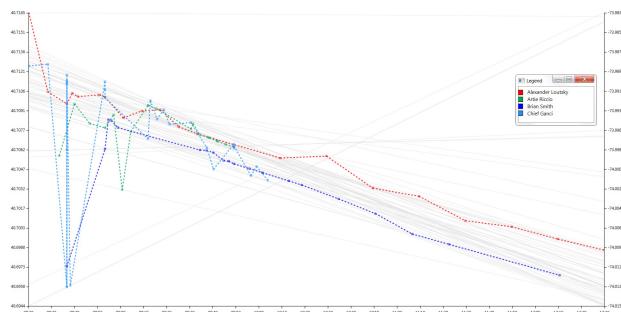


Fig. 2: Storylines of three firefighters and one EMT.

VII. Next steps: violations taxonomy and the South Africa truth and reconciliation corpus

Extensions of this work will begin with analysis of South Africa's Truth and Reconciliation Proceedings. Over a two-year period, the TRC collected 7,000 amnesty applications and 22,000 witness statements describing 34 years of abuses occurring nationwide. This material will expand our techniques to larger geographic contexts, more diverse time frames, elements of a second-language, and a much wider range of rights violations. As each mass violation event has a particular violations ecology¹³, one challenge for violations researchers is to identify the emblematic subset of violations. Currently, we are working on methods for the automatic correlation of a violation description to a taxonomy of violations¹⁴. This method will help classify the narrative events in the context of particular violations within each narrative and further our work in entity resolution. It will also expand our methods into what Salway and Herman refer to as top-down, hypothesis-driven, and bottom-up, data-driven, methods¹⁵. Additional visualization elements currently under development include UI tweaks to automatically generate a colorspace, and analytic affordances allowing for the drilling down from Storygrams to the original source document fragments.

References

1. Felman, S. (1995) *Education and crisis, or the vicissitudes of teaching*. In Trauma: Explorations in memory, C. Caruth, Ed. JHU Press, pp. 13 – 60.
2. Torchin, L. (2012) University of Minnesota Press.
3. White, H. *The structure of historical narrative* (2010). In The fiction of narrative: Essays on history, literature, and theory, 1957–2007. JHU Press.
4. Bal, M. (1997) *Narratology: Introduction to the theory of narrative*. University of Toronto Press.
5. Ball, P.D. (1996) *In Who did what to whom?: planning and implementing a large scale human rights data project*, American Association for the Advancement of Science Washington, DC.
6. Miller, B., Shresta, A., Derby, J., Olive, J., Umpathy, K., Li, F., and Zhao, Y. Z. (2013) *Digging into human rights violations: Data modelling and collective memory*. In IEEE Big Data 2013 (10 2013).
7. Aigner, W., Miksch, S., Müller, W., Schumann, H., and Tominski, C. (2007) *Visualizing time-oriented data systematic view*. Computers & Graphics 31, 3, 401–409.
8. Jern, M., and Franzen, J. (2006) "geoanalytics" - exploring spatio-temporal and multivariate data. In Proceedings of Tenth International Conference on Information Visualization (july 2006), pp. 25 –31.
9. Fry, B. (2000) *Organic information design*. Master's thesis, Massachusetts Institute of Technology, Cambridge, MA.
10. Ohno, S., Saito, S., and Inaba, M. (2010) *A platform for mining and visualizing regional collective culture*. In Culture and computing. Springer, pp. 188–199.
11. Shresta, A., Miller, B., Zhu, Y., and Zhao, Y. (2013) *Storygraph: extracting patterns from spatio-temporal data*. In Proceedings of the ACM SIGKDD Workshop on Interactive Data Exploration and Analytics (New York, NY, USA), IDEA '13, ACM, pp. 95–103.
12. Inselberg, A. (1985) *The plane with parallel coordinates*. The Visual Computer 1, 2, 69–91.
13. Cross, N., and Jarvis, H. (1999) CGDB: Cambodian Genocide Data Bases. Input Manual for CBIB: CGP Bibliographic Database. Documentation Center of Cambodia, Phnom Penh, Cambodia.
14. Dueck, J., Guzman, M., Verstappen, B., and Huridocs. (2001) *Micro-thesauri: a tool for documenting human rights violations*, vol. 12. and distributed by Huridocs Advice and Support Unit/Secretariat 48 chemin du Grand-Montfleury CH-1290 Versoix Switzerland Tel. 41.22. 755 5252, fax 41.22. 755 5260 Electronic mail: huridocs@comlink.org Website: www.huridocs.org.
15. Salway, A., And Herman, D. (2011) *Digitized corpora as theory-building resource: New methods for narrative inquiry*. In New Narratives: Stories and Storytelling in the Digital Age, R. Page and B. Thomas, Eds. U of Nebraska Press.

Clustering Search to Navigate A Case Study of the Canadian World Wide Web as a Historical Resource

Milligan, Ian

i2milligan@uwaterloo.ca
University of Waterloo

1. Introduction

From a historian's perspective, I present an approach to navigate large amounts of Internet Archive information, drawing on a case study of 4.7% of the top-level .ca domains preserved in a scrape of the entire World Wide Web, the March 2011 Wide Web Scrape. Every day, users record their thoughts, feelings, locations, ratings, votes, reviews, jokes, and so forth; an assemblage of traces of the past that historians will be able to mold into narratives. Here, I explain one way to access them beyond the WaybackMachine using open-source tools such as WARC Tools, Apache Solr, and Carrot2 Workbench.

2. Literature Review and Project Rationale

Information scholars and digital archivists are having a conversation around digital preservation and web archiving.¹ However, there is a need to approach these issues from the perspective of a historian with an interest in using web archives. My focus is on use rather than preservation.

The current way to access this material is through the WaybackMachine, run on the Internet Archive's server itself at archive.org/web or as a local installation.² The WaybackMachine is simple from a user perspective: one enters a URL and then the available dates of various snapshots are displayed across the top of the screen, and the web page is displayed if it is available.

I am concerned with the files that drive the WaybackMachine: the WebARCHive (WARC) files, the international standard for preserving Web data.³ Archiving web sites is difficult: a single page is made up of hundreds of parts, with external calls for images or other code hosted elsewhere. A WARC file provides a container for it all.

There are good reasons for a historian to be interested in these files rather than merely using the WaybackMachine. Chief among them is the latter's lack of a full-text search function. The WaybackMachine takes a URL and generates the corresponding archived website. WARC files are the system's building blocks. They contain data with can be explored another way.

3. The Canadian Internet Case Study

I draw on the 80TB Wide Web Scrape, released in October 2012 to celebrate the Internet Archive's accumulation of ten petabytes of data.⁴ The WARC files in this collection are the results of a crawl that began on 9 March 2011 and ended on 23 December 2011, totalling 2,713,676,341 websites over 29,032,069 hosts. It is important to note the limitations of this scrape. First, we do not have multiple ones: if we had more scrapes, we could compare them and thus establish temporal changes. My hope is that if we can make cogent cases for what we can do with this sort of data, more might be released. Second, the sampling practices of the Internet Archive profoundly impact this sample and are beyond my control. The exact percentage of preserved websites is unknown, but it is probably only a little more than half.⁵ Furthermore, they are not scraped and preserved on a temporally consistent basis. More work remains to be done to understand these processes, as they profoundly shape work done with it.

Beginning by downloading all 111,690 index files (one line per URL), I found the WARC files that contained the largest number of websites within the .ca domain. I subsequently downloaded the top hundred WARCs containing a total of 397,221 Canadian URLs. Based on Internet Archive statistics, this dataset is 4.7% of the indexed .ca top-level domain.

To extract information, I was primarily interested in drawing on the large body of text within this sample to analyze. Text analysis tools are most developed, and this is my active area of research interest. In order to convert these files into plain text, I relied upon the free and open-source WARC Tools collection. In short, this creates plain text files by running each website through Lynx, a text-based browser. I subsequently selected only those that had Canadian domain names, selected via regular expressions. Each of those full-text files was then transformed into an XML document with fields for the title (URL) and content (textual content of the website). This plain text conversion has also been extremely fruitful in exploring via other text analysis tools, such as Named Entity Recognition.

4. Findings

The interplay between two open-source tools, the Apache Solr search engine and the Carrot2 clustering search engine, presents a fruitful way to explore these archives. Solr is a NoSQL search engine optimized for working with millions of documents. Once data is ingested into Solr, which provides basic search, I turn to Carrot2. It is useful because of the clustering function, which takes objects and groups them into sets sharing common characteristics. While Carrot2 offers several clustering algorithms, Lingo clustering is most fruitful. Its goal is to "capture thematic threads in a search result, that is discover groups of related documents and describe the subject of these groups in a way meaningful to a human."⁶

I now want to provide an overview of what these processes were able to do in my explorations of the Wide Web Scrape. My choices of queries are necessarily limited, as this paper cannot do comprehensive justice. In my other work, I am a historian of youth cultures; how could this methodology help somebody with my research interests? Here, a query for 'children':

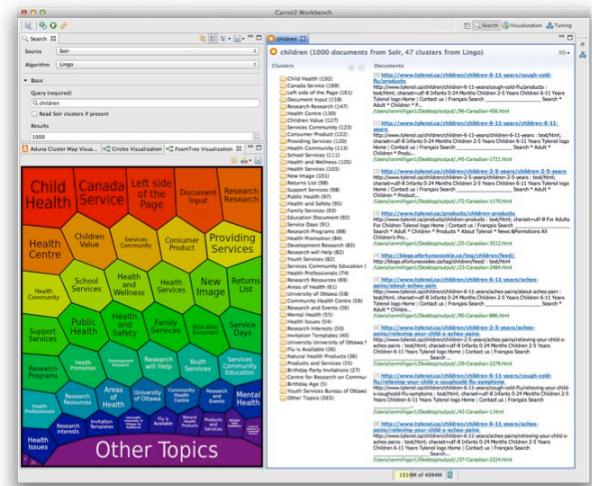


Fig. 1: The query ‘children’, demonstrating clusters within the collection

At left is an input panel, at right a list of clusters with the documents themselves. At lower left we have visualization options. This visualization is akin to an archival finding aid: we learn what these WARC files tell us about 'Children,' and whether this is worth further investigation. We see files relating to children's health (Health Canada, Tylenol), research into children at various universities, health and wellness services, as well as related topics such as Youth Services, Family Services, and mental health.

Thus, we have both an ad-hoc finding aid equivalent as well as a way to move beyond distant and close reading levels. While future studies would need to expand the amount of websites studied, we can see in this tranche of 5% of the Web that we have a good amount of information pertaining to children's health, universities, and beyond. For some researchers, this would be a boon – for others, an indication that they may need to look elsewhere. Some projects are purely exploratory, however, and with a confident sample we could begin to make overall statements concerning societal concerns towards children. Notably, this method helps us find relevant websites by putting them into easy-to-understand groups, allowing for both quantitative (number of sites) and qualitative (the sites themselves) findings.

Clusters often contain more than one object, and the relationship between clusters sheds light on the structure of a document collection. Consider the following:

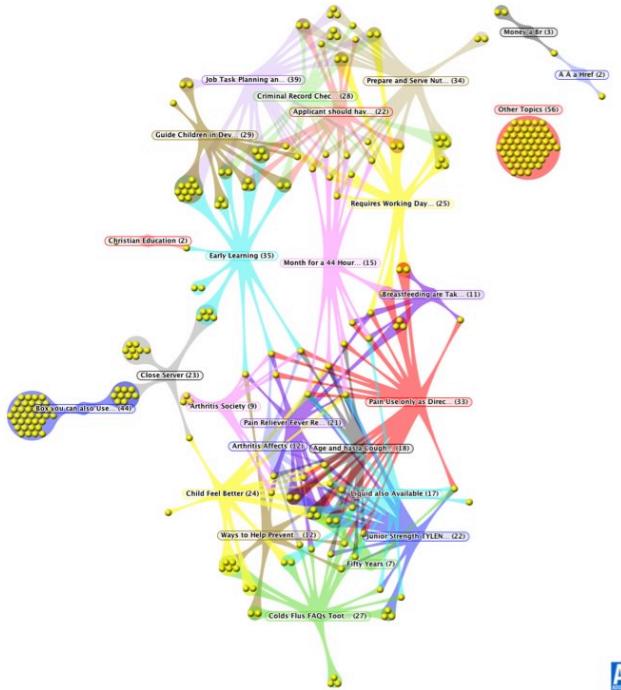


Fig. 2: The query 'children' visualized using the Aduna cluster visualization technique.

Labels represent clusters. If a document spans multiple clusters, it is represented by a dot connected to both labels, which represent clusters. For example, "Christian Education" appears in the middle left of the chart. There is one document to the left of it (partially covered by the label), a document that belongs only to it. Yet there is one to the right of it connected with "Early Learning," representing a website that falls into both categories.

From this, we can learn quite a bit about the files that we can find in the Wide Web Scrape as well as suggest which might be most fruitful for exploration. In this chart, at the bottom we see websites relating to children's health, which connect to breastfeeding, which connect to timeframes, which actually then connect to employment (which often contains quite a bit of data and time information). We then also see that connect to early childhood workers, which in turn connects to early learning more generally. The structure of the web archive relating to children reveals itself.

A downside is that the individual files are plain text. However, we can use the online WaybackMachine. Using an Automator script in OS X, a service can be configured to prefix the WaybackMachine's URL (`web.archive.org/web`) to any URL. Through three steps (service receives text, adds the prefix string, and displays the webpage), we retrieve the archived version. See Fig 3, 4, and 5 below for this process:

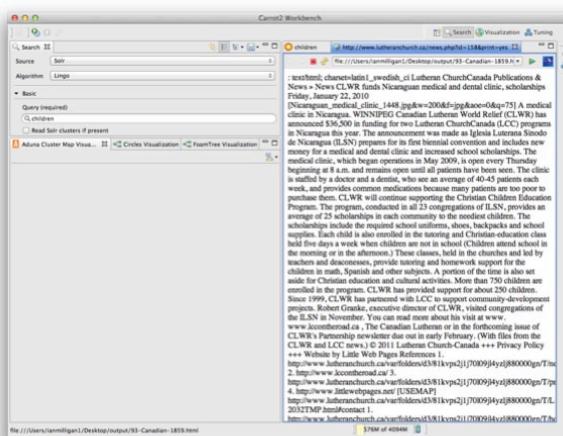


Fig. 3: An example of the plain text files that power the search engine

Fig. 4: WaybackMachine Plugin in Action

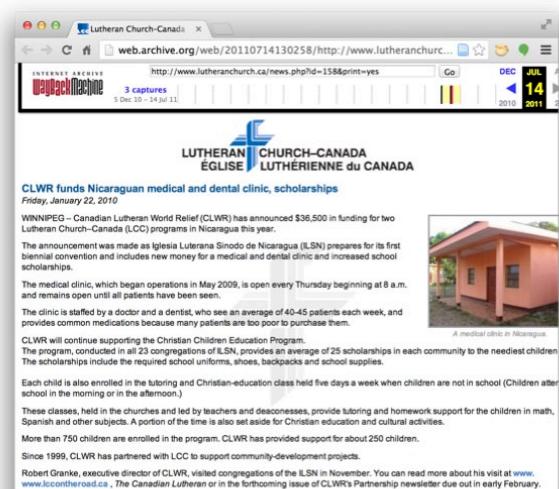


Fig. 5: From distant to close reading: the above website in the WaybackMachine.

5. Conclusions

WARC files, and web archives more generally, should be understood as key components of a future historian's professional training. Undertaking projects growing out of the 1990s or 2000s may require access to such archives. Finding aids are generally unavailable for this type of source, and would be impractical due to the sheer data quantity. This approach, integrating on-the-fly finding aid generation and access to both distant and close reading, should be considered for adoption by historians.

This project shows that by drawing on a large dataset, 4.7% of the top-level .ca domain, historians will be able to derive meaningful information and find connections between disparate bodies of information. As history enters the web era, new tools and resources will be necessary.

References

- Song, JaJa, J.** (2008). *Fast browsing of archived Web contents in 8th International Web Archiving Workshop*, Aarhus, Denmark, www.umiacs.umd.edu/publications/fast-browsing-archived-web-contents ; **Toyoda, M., Kitsuregawa, M.**, (2012). *The History of Web Archiving. Proceedings of the IEEE (Institute of Electrical and Electronics Engineers)*, 100, 1441–1443; **Brügger, N.** (2008). *The Archived Website and Website Philology: A New Type of Historical Document*. Nordicom Review, 29, 155–175; **Brügger, N.** (2009). *Website history and the website as an object of study*. *New Media and Society*, 11, 115–132; **Brügger, N.** (2010). *Web History*. Bern: Peter

Lang Publishing; **Brügger, N.** (2012). *Web History and the Web as a Historical Source*. Zeithistorische Forschungen/Studies in Contemporary History. 9; **Brügger, N.**, Finnemann, N.O. (2013). *The Web and Digital Humanities: Theoretical and Methodological Concerns*. Journal of Broadcasting and Electric Media 57, 66–80.

2. **Internet Archive** (n.d.). WaybackMachine Github. GitHub. URL github.com/internetarchive/wayback (accessed 5.29.13).

3. **ISO** (2009). ISO 28500:2009.

4. **Internet Archive** (2012). *80 terabytes of archived web crawl data available for research*. Internet Archive Blog, blog.archive.org/2012/10/26/80-terabytes-of-archived-web-crawl-data-available-for-research .

5. **Ed Summers**, *The Web as a Preservation Medium* Inkdroid.org, 26 November 2013, available online, inkdroid.org/journal/2013/11/26/the-web-as-a-preservation-medium , accessed 20 February 2014.

6. **Osiński, S., Stefanowski, J., Weiss, D** (2004). *Lingo: Search Results Clustering Algorithm Based on Singular Value Decomposition*, in: Advances in Soft Computing, Intelligent Information Processing and Web Mining, Proceedings of the International IIS: IIPWM 04 Conference. Zakopane, Poland, pp. 359–368.

The Telltale Hat: LDA and Classification Problems in a Large Folklore Corpus

Mimno, David

mimno@cornell.edu

Cornell University Department of Information Science

Broadwell, Peter M.

broadwell@library.ucla.edu

UCLA Library

Tangherlini, Timothy R.

tango@humnet.ucla.edu

UCLA Scandinavian Section and Department of Asian Languages and Cultures

Introduction

Classification is a vexing problem in folkloristics. Indexing collections that often include tens of thousands of records is essential, but neither fully manual nor fully automated methods are adequate. In this work, we combine human notions of genre and topic classification with computational classifiers and topic analysis to produce an indexing that is both appropriate for scholarly goals and robust in the presence of ambiguity.

Traditional scholarly indexes have been limited by time and technology. Although broad genre classifications such as “ballad”, “folktale”, and “legend” are well established, these formal classifications are coarse and do little more than sort the materials into large, internally diverse groupings. Most standard classification schemes assign each record to a single classification and do not allow for cross-genre classification (e.g., a ballad and a legend about the same murder will be in different categories).^{1 2 3 4} The inadequacy of these classification schemes has significantly constrained research on verbal folklore, particularly because such categorizations are often the only available topic index for any given collection.

New unsupervised machine learning methods offer scalability but lack human intelligence. Clustering algorithms partition a corpus into groups of documents that are similar. Topic modeling is more flexible, allowing each document to express multiple automatically detected themes. But such methods usually rely on simple bag-of-words representations that miss aspects of a text that are clear to readers familiar with the corpus. In addition, patterns found by algorithms may be statistically valid but uninteresting to scholars.

We explore the problem of classification in a large corpus (~35,000 records) of nineteenth-century Danish folklore and suggest possible solutions to these problems through classification and topic-modeling strategies that combine human labels with machine learning. We consider two classification schemes for the collection: in the first, each document receives one label, whereas the second assigns multiple labels to each document.

One label per story

The original collector assigned each story to exactly one of 36 labels, but we are most interested in “borderline” stories that could fit in many classes. These “liminal” stories not only reveal the challenges to classification that arise when a system can only accommodate a single label—as in the original index—but also help researchers to discover stories that are anomalous.

An excellent example of such an anomalous story appears in our target corpus, *Danske sagn* [Danish Legends]:⁵

DS_I_056:

Per Overlade was out one evening shooting hares. It was up on Kræn Møller's field. Kræn was in the process of moving his farm, and the old farm had not been completely disassembled yet, and Per intended to hide amid the old frame that was still standing and shoot a hare or two. But when he gets there, he sees an old man who is sitting in there with a red cap on who nods to him. Per gets scared and doesn't dare go in there, and so he doesn't catch any hares.

Originally labeled as a story about “mound dwellers/hidden folk,” the story could just as easily be classified in several other categories: poaching, household guardian spirits (*nisse*, suggested by the old man's red hat), and law breaking, to name but three. The story also touches on shifting agricultural practices and the significant reorganization of the Danish landscape in the early 1800s, when farms were routinely dismantled and moved out onto the newly reapportioned fields.

Where else could the editor/archivist have placed this story? To answer this question, we train a Naïve Bayes classifier by estimating a word-frequency histogram for each label. We then measure the similarity of a document to each of the resulting histograms, taking care to remove the word counts for the “query” document from the histogram for its original label. For many stories, the “true” label is the closest, but not in this case. Its top five labels in order are:

ID	Story label
36	Our forbears' way of thinking and spiritual life
35	Outdoor life
29	Witches and their sport
27	Being in league with the Devil
1	Mound dwellers/hidden folk

Although the first assignment is so broad as to be of little use—emphasizing the inadequacy of the original index—the association of the story with topic 35 highlights its affinity to stories about hunting and poaching, while topic 29 indicates the story's connection with hares—animals most commonly associated with witches.

Additionally, we can use this classification scheme to initialize a 36-topic model, creating one topic per original label. We assign each word token to the same topic as the label of its document. We then resample topic assignments for each word token in turn. Given the topic assignments of the tokens in a document, we can rank the topics for that document. After one sweep through the entire corpus, the “Mound dwellers” topic still accounts for more than 80% of the tokens in the story of Per Overlade, but after 10 sweeps, only 21% of the words remain in that topic. “Our forebears' way of thinking” and “Being in league with the Devil” instead account for a greater proportion, with the “Devil” topic triggered by words about shooting hares. Overall, the original topic class now accounts for the majority of tokens in 74% of the stories in the collection.

As we increase the number of sweeps through the corpus, the relationship between the topics of the model and the original labels becomes attenuated. At 100 sweeps, the majority of tokens remains in the original class for only 39% of the stories. In our sample story, the prominent topics are “From the time of vilenage”, “Wiverns and small creepy-crawlies”, “Our forebears’ way of thinking”, and “Death portents”. Words about shooting and hares are now assigned to the “Wiverns” topic, indicating that we should be careful in using these labels. The “Death portents” topic is represented by the words *forskrækket* (scared) and *sidder* (sitting).

Finding anomalous stories is not simply a question of precision and recall: the very fact that a story is “missed” in a given classification makes it particularly interesting. One of the jobs of the folklorist is to reconstruct the imaginary boundaries of the belief world, so stories that question or test those boundaries are the ones that are most important. Computationally cross-validating a traditional human-generated index, as described above, is an effective way to discover such liminal cases.

Multiple human-generated labels

We can also construct computational story classifiers when editors assign more than one label to each document. Human experts have catalogued a subset of the documents in our target corpus by assigning multiple labels to each document from a modern ontology that includes aspects of stories such as people, locations, and events. We would like to know how these labels map to the words in the documents, but simply counting the words in every document assigned to a label may result in noisy histograms. To improve our ability to interpret the results, we use a labeled topic model to learn which words are associated with which labels.

Multiple labels add complexity but allow us to make stronger assumptions. Since each document has more than one label, we cannot easily translate these labels into word-level assignments as in the previous experiment. On the other hand, we can be reasonably certain that the absence of a label implies that it is not relevant. Similar to LabeledLDA⁶, we can therefore estimate word-topic assignments under the constraint that words can only be assigned to one of the labels for the document, or to a “Background” label that can absorb frequent words not related to any label. We then re-estimate topic-word distributions given these assignments, and repeat the process as needed.

To evaluate the resulting word distributions, the original creator of the ontology marked individual words that are highly relevant to each label. At each stage of the algorithm, we have a ranked list of words for each label. Given relevance assignments, we can compute mean average precision (MAP) for the model at each stage. Under the initial noisy distributions, MAP for precision up to rank 20 is .26. After the first iteration, MAP increases to .33, but then begins falling in subsequent iterations, indicating that the model may be overfitting.

Consistent differences in ranking quality provide insight into labels. We are more successful at finding words related to concrete themes such as people, animals, and objects. More abstract labels, such as story resolutions and actions or events, were mostly unsuccessful. But there are exceptions: we identified no words related to the label “Farmer”, despite the fact that this is a very common label, while events such as “Disease” and “Death” identified many specific words.

Conclusion

We demonstrate that classification and topic modeling methods can be used to improve existing manual annotations in a collection of Danish folklore. We find that incorporating human labels into machine learning methods—even when the labels are noisy or incomplete—produces indexes that have the benefits of both scholarly domain expertise and data-driven analysis. We believe that these results are applicable for many corpora both in digital humanities and the wider document analysis community.

References

1. **Uther, Hans-Jörg.** (2004). *The Types of International Folktales: A Classification and Bibliography, Based on the System of Antti Aarne and Stith Thompson*. FF Communications. Helsinki: Suomalainen Tiedeakatemia.
2. **Grundtvig, Svend, Axel Olrik, Hakon Grüner-Nielsen, Karl-Ivar Hildeman, Erik Dal, Iørn Piø, Thorkild Knudsen, Svend Nielsen, and Nils Schiørring**, eds. 1966–1976 [1853–1976]. *Danmarks gamle Folkeviser*. 12 volumes. Copenhagen: Universitets-Jubilæets Danske Samfund (Akademisk forlag).
3. **Taylor, Archer.** (1934). *An Index to "The Proverb"*. FF Communications 113. Helsinki: Suomalainen Tiedeakatemia, 1934.
4. **Christiansen, Reidar T.** (1958). *The Migratory Legends*. FF Communications 175. Helsinki: Suomalainen Tiedeakatemia.
5. **Kristensen, Evald Tang.** (1892). *Danske sagn, som de har lydt i folkemunde*. Århus and Silkeborg: Århus Folkeblads Bogtrykkeri.
6. **Ramage, Daniel, David Hall, Ramesh Nallapati, and Christopher D. Manning.** (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. 248–256.

Seeing the Trees & Understanding the Forest

Montague, John Joseph

jmontagu@ualberta.ca
University of Alberta, Canada

Rockwell, Geoffrey

University of Alberta, Canada

Ruecker, Stan

IIT - Institute of Design, USA

Sinclair, Stéfan

McGill University, Canada

Brown, Susan

University of Alberta, Canada

Chartier, Ryan

University of Alberta, Canada

Frizzera, Luciano

University of Alberta, Canada

Simpson, John

University of Alberta, Canada

“Humanists have always been explorers. They sail not on seas of water but on seas of color, sound, and, most especially, words.” — John B. Smith, 1984

Why is big data such a big deal? Modern digital communications are producing and recording data at a feverish rate, with 90% of the world’s recorded data, most of it unstructured, having been produced in just the last two years (Dragland, 2013). In addition, more traditional texts are being digitized all the time, and Crane (2006) tells us that while the largest academic digital libraries now hold tens of thousands of books, a completed Google Library will likely have more than *ten million*. To further Smith’s analogy, we are adrift in a sea of data, and we are at risk of floundering.

Front page news events like Edward Snowden’s May 2013 revelations disclosing the secret NSA (America’s National Security Agency) collection and analysis of massive amounts of ostensibly private data both domestically and internationally, are making big data analytics part of the public lexicon. Major corporations, led by IBM with \$1.3 billion in big data revenue in 2012, are scrambling to adopt practices that will allow them to capitalize on the volume of information now being generated.

Global big data related revenues in 2012 were \$11.6 billion, and are projected to break \$18 billion in 2013 (Kelly et al.).

While one might shudder to think what the NSA will do with all that information, outside of the world of politics and espionage, what can researchers and academics do with the volume of information now, or at least soon to be at our disposal? Matthew Jockers asks, "How do we mine them [texts] to find something we don't already know?" (University of Nebraska-Lincoln) To utilize such massive corpora and avoid succumbing to the dilemma of what McCormick et al (1987) refer to as "information without interpretation" we need to find the most effective ways for end-users to explore, visualize and interact with big data such that it is both meaningful and understandable, if possible by both the trained and untrained eye.

Big data analytics and computer-supported visualization offer ways to read collections as our cognitive abilities are stretched to their limit with the sheer volume of data available (Araya, 2003). However, as Franco Moretti has pointed out, what we are reading when we use text mining methods and visualizations are really models of the collections. (Moretti 2013, p. 157) It is important therefore to survey the visual models emerging and question if they can better be designed to suit humanities exploration. In this paper we therefore propose to look at visualization for text mining in the following ways:

- **Survey** text mining visualization in the humanities. Who is using text mining and why? What kinds of visualizations do they find compelling and why?
- **Identify and Combine** commonly presented visualizations for modeling. What visual models could be used for the exploration of large corpora? How could they be combined?
- **Model** interactive prototypes of different combinations of visualizations for exploration.

Surveying: In a poster at DH 2013 we presented a framework of text mining tools that are useful in the humanities. Now we will survey the variety of visualizations used to present mining results. We will begin with early discussions of visualization like Smith's "Computer Criticism" (Style 1978, p. 326), where he notes with agreement Paul de Man's observation that as late as 1973 there had been no evolution beyond the close reading "techniques of description and interpretation" being used by literary critics since the 1930s or 40s, and suggests "pictorial representation" as one of the potential uses of computer aided text analysis. Brunet in a 1989 article talks about exploiting large corpora and provides a number of examples of visualizations. Our survey will examine advances in practice and understanding in the intervening thirty plus years, leading to contemporary works like Franco Moretti's *Graphs, Maps, and Trees*, which, as the title suggests, introduces visual models for literary exploration.

Our survey pays particular attention to recent text mining projects and tools including "The Proceedings of the Old Bailey", David L. Hoover's work with cluster analyses at NYU using "MiniTab", Matt Jockers' topic modeling work in "Macroanalysis", the University of Waikato's "WEKA", the open-source visualization tool "Gephi", the UMass machine learning tool "MALLET", the research tools for textual study reviewed on "TAPoR" as well as some of our own INKE related projects such as "Dynamic Table of Contents", "CiteLens", "TextTiles" and "dialR".

Identifying and Combining: While we recognize that output will assume a variety of formats including for instance heat maps, topographic plots or scatter plots, our survey suggests that text-mining projects commonly use five principal types of visualization:

- 1. *Dendograms*, showing data clustering within sets
- 2. *Histograms*, showing change over time
- 3. *Network diagrams*, showing how entities are connected within a network
- 4. *Word clouds*, representing topics of words
- 5. *Scatter plots*, showing words or parts in an abstract space

Visualizations that are presented in print are typically Spartan, focusing the attention on the results through careful design. All affordances are removed. The same types of visualizations automatically generated from large data sets, however, tend to be too dense to be useful and have to include

affordances if meant to be interactive. Simple representative visualizations, like histograms for instance, are insufficient to display complex interrelationships. Dendograms, especially if you are working with massive data sets, quickly become an illegible mass of inter-connectivity. Word clouds are good at showing the relative frequency of words in a text or topic, but not at comparing one text or topic to another. Network diagrams produce some beautiful results, but suffer from the same difficulties as dendograms; large data sets quickly lead to illegibility.

"By visualizing information, we turn it into a landscape that you can explore with your eyes, a sort of information map. And when you're lost in information, an information map is kind of useful." — David McCandless, 2010

Modeling: Our research now is looking at ways to increase the exploratory power of visualization of large data sets. Used in combination, visualizations can provide otherwise elusive insight and clarity. They can also provide affordances for each other – a histogram can be used to explore a dendrogram. We have developed interactive prototypes using combinations of visualizations, and focusing on not simply allowing, but even *encouraging* the user to truly explore the (big) data, "function[ing] almost instinctively", as McCullough (1996) stated, "to serve the process of development". The more we can encourage users to explore and play with the data, the more likely they are to develop useful insights.

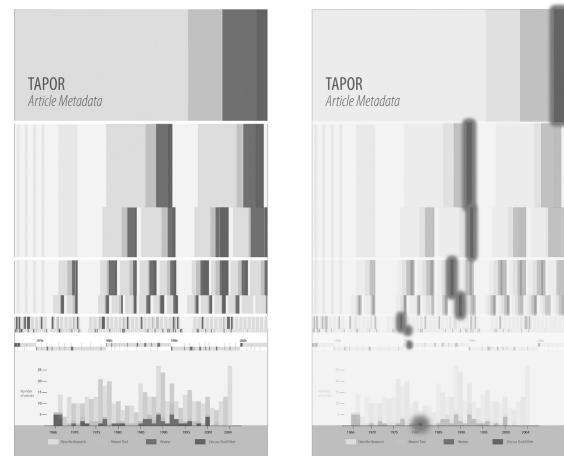


Fig. 1: Combination of a modified dendrogram showing clustering, and a diachronic timeline of 500 philosophy papers.

Figure 1 shows a prototype of a combination of dendrogram and histogram that we developed to visualize the clustering of 500 papers published between 1966 and 2004. The prototype is a combination of a modified dendrogram showing the clusters, and a diachronic visualization displaying the subjects over time. The displayed data is intuitively explorable, and the modified dendrogram is designed to encourage exploration. We developed the R code to prepare the data for interactivity. In Voyant we have developed skins that combine scatter plots and histograms, word clouds and histograms, and network diagrams with other tools. Again, the tools are open for others to recombine.

To conclude, surveying commonly used graphical representations allowed us to identify commonly used visualizations that humanists find useful. To scale these so that they can be used interactively to explore large data sets we have prototyped combinations, where one visualization can be used to explore another and vice versa. The goal is visualizations that help researchers make sense of big data; visualizations that let us explore the forest, not just the trees so that we can draw accurate and appropriate inferences from the data.

References

- The Proceedings of the Old Bailey* - <http://www.oldbaileyonline.org/>
- Cluster Analysis, Principal Components Analysis (PCA), and T-testing in Minitab* - <https://files.nyu.edu/dh3/public/ClusterAnalysis-PCA-T-testingInMinitab.html>
- Jockers, M.** *Macroanalysis*.
Weka 3: Data Mining Software in Java - <http://www.cs.waikato.ac.nz/ml/weka/>
- Gephi - The Open Graph Viz Platform* - <https://gephi.org/>
- Mallet - MACHine Learning for LanguagE Toolkit* - <http://mallet.cs.umass.edu/>
- TAPoR - Research Tools for Textual Study* - <http://tapor.ca/>
- Dynamic Table of Contents* - http://www.ualbertaprojects.info/dyntoc/dyntoc_v3_5/Main.html
- CiteLens* - <http://labs.fluxo.art.br/CiteLens/>
- TextTiles* - <http://dev.giacometti.me/textTiles/trunk/>
- dialR* - <http://research.artsrn.ualberta.ca/~dialr/drMain.html>
- Allison, S., Heuser, R., Jockers, M., Moretti, F.** and **Witmore, M.** (2012). *Quantitative Formalism: An Experiment*. Pamphlet 1, Stanford Literary Lab. Print.
- Araya, A. A.** (2003). *The Hidden Side of Visualization*. Techné: Research in Philosophy and Technology; Vol 7, No 2, Print.
- Brunet, É.** (1989). *L'Exploitation des Grands Corpus: Le Bestiaire de la Littérature Française*. Literary and Linguistic Computing, Vol. 4, No. 2, p. 121-134. Print.
- Crane, G.** (2006). *What Do You Do with a Million Books?* D-Lib Magazine Vol. 12 No. 3, Print.
- Dragland, A.** (2013). *Big Data, for better or worse*. SINTEF.no. 22 May 2013. Web. 27 Oct. 2013.
- Jockers, M.** (2013). *Macroanalysis : digital methods and literary history* - University of Illinois Press, Urbana, Chicago & Springfield.
- Kelly, J., Foyer, D., Vellante, D. and Miniman, S.** (2013). *Big Data Vendor Revenue and Market Forecast 2012-2017*. Wikibon.org. 19 Feb. 2013. Web. 28 Oct. 2013.
- McCandless, D.** (2010). *David McCandless: The beauty of data visualization*. TED Talks. Web Video. 25 Oct. 2013.
- McCormick, Bruce H., DeFanti, Thomas A., and Brown, Maxine D.** (1987). *Visualization in Scientific Computing*. Computer Graphics 21, 6 (November). New York: Association for Computing Machinery, SIGGRAPH, Print.
- McCullough, M.** (1996). *Abstracting Craft: The Practiced Digital Hand*; Cambridge, MIT Press, Print.
- Moretti, F.** (2013). *The End of the Beginning: A Reply to Christopher Prendergast*. Distant Reading. London: Verso, p. 137-158. Print.
- Risen, J. and Poitras, L.** (2013). *NSA Gathers Data on Social Connections of U.S. Citizens*, New York Times, 28 Sep. 2013. Web, 27 Oct, 2013.
- Simpson J.; Rockwell, G.; Sinclair, S.; Uszkalo, K.; Brown, S.; Dyrbye, A.; Chartier, R..** (2013). *Framework for Testing Text Analysis and Mining Tools*. Poster presented at the Digital Humanities 2013 conference at the University of Nebraska-Lincoln. Lincoln, Nebraska, USA.
- Smith, J.** (1978). *Computer Criticism*. Style. Vol XII, No 4. Print.
- Smith, J.** (1984). *A New Environment For Literary Analysis*. Perspectives in Computing 4. 2/3, (1984): 20-31. Print.
- Tufte, Edward.** (1983). *The Visual Display of Quantitative Information*; Cheshire, CT: Graphics Press, Print.
- University of Nebraska-Lincoln.** (2012). *By text-mining the classics, UNL prof uncovers new literary insights*. UNL News Blog. 23 Aug. 2012. Web. 27 Oct. 2013.

Making Digital Humanities Work

- Munoz, Trevor**
trevor.munoz@gmail.com
 University of Maryland
- Guiliano, Jennifer**
jenguiliano@gmail.com
 University of Maryland

Abstract

In the many conversations touched off by Alan Liu's question, "Where is cultural criticism in the digital humanities?", there has been little attention paid to the first term of that question: *where*? Treating this term abstractly has robbed many of the resulting discussions of the relationship between the digital humanities and (what Liu terms) "the mainstream humanities" of their usefulness. We argue that focusing precisely on the where of the digital humanities, that is on sets of material practices employed at a specific site during the production of digital humanities work opens stale questions to new analytical approaches and, much better, to new avenues of intervention.

Liu's Description of Digital Humanities as Caricature

Most responses to Liu's question have pursued answers in terms of disciplinary identity and its accompanying political and organizational boundaries (intellectual history honed to an instrumentalist point)¹. In the version of his essay on this question published in the Debates in the Digital Humanities volume, Liu himself focuses on the production and consumption of "the digital humanities" in terms of what he calls "the great postindustrial, neoliberal, corporate, and global flows of information-cum-capital"². This commitment to examining the digital humanities from a world system view does much to explain, if not excuse, the caricature of digital humanities work that follows: "It is as if, when the order comes down from the funding agencies, university administrations, and other bodies ... digital humanists just concentrate on pushing the "execute" button...." From previous work such as The Laws of Cool³ through "Where is Cultural Criticism in the Digital Humanities?" and forward into essays such as "The Meaning of the Digital Humanities"⁴, Liu's critical work has been intermixed with advocacy for a vision of the humanities that emphasizes resistance to an academic culture that is "postindustrial, neoliberal, corporate", etc. Thus it would seem a willful misreading to understand Liu's portrayal of digital humanities work in the Debates essay as something other than a continuation of his advocacy for a particular regime within the world system he constructs. Yet, many readers seem to have taken this figuration of digital humanities as descriptive rather than rhetorical.

Work Sites as Opportunities for Analysis

Indeed, to save the question of "Where is cultural criticism in the digital humanities?" from tendentiousness requires richer, more detailed accounts of the digital humanities as sets of material work practices⁵. As Clifford Geertz complained of structuralist accounts in anthropology, "Whatever, or wherever, symbol systems 'in their own terms' may be, we gain empirical access to them by inspecting events, not by arranging abstracted entities into unified patterns"⁶. As a contribution to the debate over the relationship of the digital humanities to the humanities, we offer an account of attending to, and then trying to re-make, specific material practices at a site of digital humanities work. The specific case we will discuss involves challenges that have arisen from a model of organizing work around "fellowships", a structure inherited and adapted from the "mainstream humanities". We will narrate the creation and formulation of a model for organizing work practices created in response to perceived structural flaws in the fellowship model. This "anti-fellowship" model, which we have called "The Digital Humanities Incubator" (henceforth, the Incubator) prizes long-term enrichment by the variety of participants engaged in DH work versus short-term gain by project-driven "fellows." Our discussion of the Incubator then is a critical reflection on the the program and our thought process in revising it from one iteration to the next. The Incubator functions as a work of critical design⁷; it is one of our tools for thinking through how to make digital humanities work.

This example is drawn from the site we are best able to describe in sufficient detail, our own place of work, the Maryland Institute for Technology in the Humanities (MITH), an active digital humanities center in the U.S. Of course "digital humanities" is a purposefully capacious term used to cover many different types of work by many different practitioners⁸. Furthermore, we acknowledge that there are relatively few centers and therefore relatively few digital humanists who work in them relative to the scale of the global digital humanities field. Yet, as Donna Haraway argues in her essay on situated knowledges, the goal should not be "a doctrine of objectivity that promises transcendence, a story that loses track of its mediations just where someone might be held responsible for something, and unlimited instrumental power"⁹. We certainly do not mean to advance any singular account of digital humanities work practices nor do we make claim to "unlimited instrumental power" for our descriptions. Yet, it is our hope by focusing on the practice of DH labor related to fellowships that we can encourage practitioners to explore the material realities of their own practice.

Organizing Work: A Brief History of the Fellowship Model

Popularized by the Institute for Advanced Technology in the Humanities (IATH) at the University of Virginia and adapted at MITH and other digital humanities sites (not all officially centers), digital humanities fellowships have served to connect faculty—in established and ostensibly fixed and stable organizational roles—to newer digital humanities activities. At the same time, these fix these novel, alternative and contingent digital activities to existing structures under the rubric of innovation or enhancement to scholarly activities. Fellowship experiences were most often configured as short-term (1-2 years), focused on the creation of a specific product, tool, project, and reliant divisions of labor between participants with unequal agency and power¹⁰. Faculty could be recruited as participants, supporters, and advocates and the digital humanities initiative could demonstrate its effect on the local academic community—as service, if nothing else. For a new endeavor in the fundamentally conservative academy, the adoption of a pre-made structure of organizing labor was politically astute and laid much of the groundwork for the wide success of the field today. Fellows became advocates for the digital humanities and examples of its proliferation into humanities disciplines. Yet it is crucial to remember that this model—leave from the assumed structures and schedules of faculty life, provision of technology, and access to staff with computational or other forms of alternative expertise—is an artifact of a particular historical moment. By historicizing “the fellowship” we mean to disrupt a tendency toward a teleological or developmental view of this type of work organization as a necessary step in the introduction of the digital humanities to a new site. We do not claim that these fellowship arrangements have not produced some excellent digital humanities scholarship. However, based on our experiences working in and managing these activities, we began to feel that the arrangement suffered from flaws not in just its execution in any particular case but structurally overall.

What Studying Work Can Do

The distinction between individual project successes and structural flaws is important to emphasize because it contains the most crucial lessons to be learned for the digital humanities and humanities enterprise. By shifting focus from particular work products or outputs (where Liu and many of his respondents focus their attention) to the structuring of material conditions at the site of work production, which make certain kinds of digital humanities possible, we engage issues raised by Richard Sennett's notion of "the workshop" as "a productive space in which people deal face-to-face with issues of authority"¹¹. Sennett observes that "the successful workshop will establish legitimate authority in the flesh [in the form of

recognized mastery of skill], not in rights or duties set down on paper." In too many cases, a faculty fellow is not the head of a happy, successful workshop in these terms even if the end product is lauded. Using ethnographic approaches to the study of work, especially as filtered through fields such as science and technology studies (an avenue suggested by Liu himself), perspectives from critical and participatory design^{12 13}, feminist theory, and labor studies allows us to unpack the specific material practices of labor in the contemporary digital humanities project-as-work-site/workshop. As careful articulations of the material conditions and practices of a particular site of work intersect with questions of power and authority, the management of digital humanities work can provide productive interventions in questions such as the disciplinary identity of the humanities and, implicitly, their future.

References

1. **Ramsay, Stephen.** (2013). *Why I'm In It*. Stephen Ramsay. https://web.archive.org/web/20131026120131/stephenramsay.us/2013/09/12/why_im_in_it/.
2. **Liu, Alan.** (2012). *Where Is Cultural Criticism in the Digital Humanities*. In *Debates in the Digital Humanities*, edited by Matthew K. Gold. Minneapolis: University of Minnesota Press.
3. **Liu, Alan.** (2004). *The Laws of Cool Knowledge Work and the Culture of Information*. Chicago: University of Chicago Press.
4. **Liu, Alan.** (2013). *The Meaning of the Digital Humanities*. PMLA 128 (2): 409–23.
5. **Latour, Bruno, and Steve Woolgar.** (1986). *Laboratory Life: The Construction of Scientific Facts*. Princeton, N.J.: Princeton University Press.
6. **Geertz, Clifford.** (1973). *The Interpretation of Cultures*. [New York (N.Y.)]: Basic Books.
7. **Dunne, Anthony, and Fiona Raby.** (2013). *Speculative Everything: Design, Fiction, and Social Dreaming*.
8. **Kirschenbaum, Matthew G.** (2010). *What Is Digital Humanities and What's It Doing in English Departments?* ADE Bulletin, no. Number 150: 1–7.
9. **Haraway, Donna.** (1988). *Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective*. Feminist Studies 14 (3): 575. doi:10.2307/3178066
10. **Bradley, John.** (2009). *What the Developer Saw: An Outsider's View of Annotation, Interpretation and Scholarship*. Digital Studies / Le Champ Numérique 1 (1). www.digitalstudies.org/ojs/index.php/digital_studies/article/view/143.
11. **Sennett, Richard.** (2008). *The Craftsman*. New Haven: Yale University Press.
12. **Björgvinsson, Erling, Pelle Ehn, and Per-Anders Hillgren.** (2010). *Participatory Design and 'Democratizing Innovation'*. In Proceedings of the 11th Biennial Participatory Design Conference, 41–50. PDC '10. New York, NY, USA: ACM. doi:10.1145/1900441.1900448. doi.acm.org/10.1145/1900441.1900448.
13. **Dantec, Christopher A. Le, and Carl DiSalvo.** (2013). *Infrastructuring and the Formation of Publics in Participatory Design*. Social Studies of Science 43 (2): 241–64. doi:10.1177/0306312712471581.
14. **Balsamo, Anne Marie.** (2011). *Designing Culture: The Technological Imagination at Work*. Durham [NC]: Duke University Press.

Tracking Semantic Drift in Ancient Languages: The Bible as Exemplar and Test Case

Munson, Matthew

mmunson@gcdh.de
Georg-August-Universität Göttingen

Language changes. On this everyone would agree. But how can we track this ever-changing phenomenon? If we focus on modern languages, the task is easier since we have native speakers whom we can ask, "How is this usage different than this other one?" But in the case of historical languages, and especially those spoken and written millennia ago, this task becomes much more difficult. How is it possible for us to create, or at least simulate, in ourselves the language proficiency of a society that has been dead for hundreds or thousands of years? And if we cannot rely on native proficiency, how can we track systematic language change and, thus, come to a better understanding of the language and texts of any particular period. The pioneering work most closely associated with John Sinclair gives us our best answer: Trust the Text!¹ We have millions of words of, e.g., Greek, ranging over a time-span of 3000 years from Homer to the present day.² What we need are methods that can help us to harness this huge amount of information. David Bamman and Gregory Crane have already begun working in the field of historical word-sense variation in Latin.³ Relying on translation equivalents, they were able to successfully track word sense variation in Latin in a 389-million word corpus. By their own admission, however, this method has the drawback of requiring "large amounts of parallel text data"⁴ in translation. In contrast, the method proposed here, comparison of co-occurrence patterns in two or more corpora, which has been applied in many other fields (see below), only requires simple, plain-text input in a single language.

The first theoretical foray into computational analysis of co-occurrence patterns came in 1955 with Warren Weaver's article "Translation."⁵ Starting from the recognition that the sense of any word is ambiguous if examined in isolation, he asserts, "But if one lengthens the slit in the opaque mask, until one can see not only the central word in question, but also say N words on either side, then if N is large enough one can unambiguously decide the meaning of the central word."⁶ The necessary corpora and computational power to realize Weaver's theory, however, only came much later.

Computational analysis of co-occurrence patterns on large-scale corpora began with the COBUILD project, which set out to build “the very first dictionaries to be based completely on corpus data” and, in doing so, systematically tracked collocations, defined as “the high-frequency words native English speakers naturally use with the target word.”⁷ Since then, co-occurrence analysis has been used in several fields in which word-sense disambiguation is necessary, such as speech recognition,⁸ machine translation,⁹ and topic modeling,¹⁰ and it is the basis for the field of distributional semantics.¹¹

In this paper, I will present my application of co-occurrence analysis to the problem of historical word-sense variation, what I call in my title “semantic drift.” I have chosen to carry out these experiments using as my two corpora the Greek Old Testament (the Septuagint) and the Greek New Testament for several reasons: the texts are easily available in digital form, have been deeply researched and, thus, deeply annotated, exist in multiple translations that can be used to benchmark methods and to test results, are of great interest to millions of people around the world, and, finally, because they are the most influential texts in the history of western civilization, the research can be easily extended to other corpora and, with more difficulty, even into other contemporary languages such as Latin, Hebrew, Aramaic, and Coptic, to name just a few. The presentation will have two primary foci: the method and exemplary results.

The method consists of the following steps. First, I tokenized the texts and calculated co-occurrence counts for every word in an 8-word window.¹² Using these co-occurrence tables, I calculated the statistical significance of each collocate word to each node word using the log-likelihood measure as described by Manning and Schütze.¹³ Log-likelihood was chosen primarily because it deals very well with sparse data and can be easily interpreted without recourse to, e.g., chi-squared tables.¹⁴ The former is important because most data in language is quite sparse and I was reluctant to eliminate a large amount of my data simply because the chosen method could not deal well with it.¹⁵ Ease of interpretation was important because, instead of using the measure as a means of hypothesis testing, in which

I would expect to get a yes or no answer, I used it as hypothesis weighting, i.e., to measure how much more likely one thing is than another. My purpose is not to decide if two words certainly form a set collocation but, instead, to measure the strength of collocation, ranging from strong repulsion to strong attraction, and compare this range with the ranges of other node words to find relationships. Having calculated the statistical significance of these relationships, I used the cosine similarity¹⁶ measure to determine the strength of relationship, first, of every word in the Old Testament with its counterpart in the New Testament (e.g., Θεός (God) in the Old Testament to Θεός in the New Testament) and, second, of every word in the Old Testament with every other word in the Old Testament and the same for the New Testament. These results allow me to discover which words' senses have changed the most (comparison of Old Testament to New Testament) and how they have changed (comparison of the words most similar to, e.g., Θεός in the Old Testament with those most similar to Θεός in the New Testament).

Old Testament		New Testament	
Σαλωμών	Solomon	Σατανᾶς	Satan
εξολοθρεύω	destroy utterly	Φαρισαῖος	Pharisees
πρόσταγμα	ordinance, command	μέλλω	to be about to, destined
παιδίον	little boy, child	πιστεύω	to trust, believe
πεδίον	plain, field	ἐκαποντάρχης	leader of a hundred, centurion
αἰχμαλωσία	captivity	ὑπάμενος	to go (away), die
ἔγκαταλείπω	leave behind, desert, forsake	τοιοῦτος	like this, certain
ἄλλοφυλος	alien, foreign	πίστις	trust, faith
κτήνος	domestic animals, cattle	εὐαγγέλιον	good news, gospel
έναντινος	opposite, facing	Πέτρος	Peter
ἔκδητέω	seek out, search for	Πιλάτος	Pilate
χριστός	anointed, christ	Χριστός	Christ
ἔλπιζω	hope	οὖν	and so, therefore
καταλαβάνω	to take, lay hold of	χάρις	grace, gift
ἔλαυω	to lift up, drive away	ἵνα	in order to
Βεναγιάν	Benjamin	φανερώδως	reveal, show
έναντι	opposite, before	ἀπόστολος	apostle
συντελέω	to finish	ἀγάπη	love
κακία	wickedness	Παύλος	Paul
πατέρουσ	to strike, smite	μαθητής	disciple

Fig. 1: Results based on the differences in cosine similarity measure between Θεός (God) and the list words. Those on the left are nearer to Θεός in the OT, on the right to Θεός in the NT.

After relying purely on computational methods to this point, the final results of my research come through qualitative analysis of the comparisons described in the previous paragraph. The two tables above show the 20 words most closely associated with Θεός (God) in the Old Testament and the New Testament based on the differences between the cosine similarity scores in each testament between Θεός (God) and the words in the list. The colors have been added by me to highlight what I see to be related words in each list. What we see on the left is that God in the Old Testament is more closely related to words concerning ruling (in yellow: "Solomon", "command", "anointed", "Benjamin"), violence (red: "destroy utterly", "to lay hold of", "to drive away", "to strike"), agriculture (brown: "field", "cattle"), and the Exodus (green: "captivity", "foreign"). While in the New Testament, God is more closely related to (evil) rulers (yellow: "Satan", "Pharisees", "centurion", "Pilate"), servants of God (dark purple: "Peter", "Christ", "apostle", "Paul", "disciple"), and words that relate the servants to God (light purple: "to believe", "faith", "gospel", "grace", "love"). So, by classifying the words most closely related with Θεός (God) in each of the testaments, we are able to determine not only that the portrayal of God had changed from the Old Testament to the New Testament, but also to see how it changed (move from a ruler who leads and makes war to a patron who offers to and receives favors from clients) and to guess at the probable historical cause (change from an independent monarchy to a Roman province). In the second part of this paper, salient examples, such as that described above for God, will be used to demonstrate the effectiveness of this method.

The final section of the paper will be a look forward at how this method could be extended to other corpora and even other languages, allowing us to tell the stories of language development with more precision and so, ultimately to understand historical texts better.

References

1. John McHardy Sinclair (2004). *Trust the text : language, corpus and discourse*. London: Routledge.
2. The Thesaurus Lingua Graece collection claims to contain 105 million words of Greek "from Homer (8 c. B.C.) to the fall of Byzantium in AD 1453 and beyond." Thesaurus Lingua Graece. n.d. 1 November 2013. www.tlg.uci.edu.
3. David Bamman and Gregory Crane (2011). *Measuring Historical Word Sense Variation*. www.perseus.tufts.edu/publications/bamman-11.pdf . 30 October 2013.
4. Bamman and Crane, 1.
5. Warren Weaver (1955). *Translation*. www.mt-archive.info/Weaver-1949.pdf . 30 October 2013.
6. Weaver, 8.
7. www.mycobuild.com/about-cobuild.aspx
8. www.mycobuild.com/about-cobuild.aspx
9. en.wikipedia.org/wiki/Machine_translation#Disambiguation
10. Mark Steyvers and Tom Griffiths (2007). *Probabilistic Topic Models*. psiexp.ss.uci.edu/research/papers/SteyversGriffithsLSABookFormatted.pdf . 30 October 2013.
11. This field was pioneered especially by J.R. Firth and Zellig Harris in the 1950s. See especially Zellig S. Harris, „How Words Carry Meaning.“ 1986. *Language and Information: The Bampton Lectures*, Columbia University, 1986. Lecture.
1. November 2013. www.ircs.upenn.edu/zellig/3_2.mp3 and John Rupert Firth. "A synopsis of linguistic theory 1930-1955." *Selected Papers of J.R. Firth, 1952-1959*. Ed. F.R. Palmer. Harlow: Longmans, 1968. 168-205.
12. Léon, p. 14, footnote 15.
13. Christopher Manning and Hinrich Schütze (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts: MIT Press. 172-175.
- Manning and Schütze, 172. On the topic of log-likelihood and spare data, see Ted Dunning. „Accurate Methods for the Statistics of Surprise and Coincidence.“ March 1993. *ACL Anthology: A Digital Archive of Research Papers in Computational Linguistics*. 1. November 2013. acl.ldc.upenn.edu/J/J93/J93-1003.pdf
15. Dunning, 61.
16. Manning and Schütze, 299-303.

Bridging the Local and the Global in DH: A Case Study in Japan

Nagasaki, Kiyonori

nagasaki@dhii.jp

International Institute for Digital Humanities

Muller, A. Charles

Graduate school of Humanities and Sociology, University of Tokyo

Tomabechi, Toru

International Institute for Digital Humanities

Shimoda, Masahiro

Graduate school of Humanities and Sociology, University of Tokyo

As the term "Digital Humanities" has been gradually gaining attention around the world, with researchers in English-speaking countries gathering under this banner in increasing numbers. Among these, there are the earlier scholars who had previously known the field as Humanities Computing; and there are also scholars who have become involved more recently, directly under the rubric of Digital Humanities. Yet another new trend is that where scholars from non-English-speaking and non-Western countries have also been gradually getting involved in the international DH community. One of these recent entrants to the international DH community is the Japanese Association for Digital Humanities (JADH). The presence of this new organization constitutes one piece of evidence to show that the international community is gradually broadening the scope of its membership. This trend has been actively supported by the Multi-lingualism and Multi-culturalism Committee of the ADHO, as well as by individual scholars who believe forming

a global community can only enrich DH and the humanities. In view of this fact, it seems that it will become worthwhile to release the CFP of the DH conference in many languages.

Additionally, recently several non-English Western communities have been established. For example, Hispanic, Italian, and German DH were discussed during the DH2012 conference at Hamburg. Each language area has long and deep history to engage in the research and practice of DH. Moreover, Global Outlook::Digital Humanities (GO::DH) has started to cover a wider area, such as Latin America, China, Africa, and so on, especially focusing on communication and collaboration across and between High, Mid, and Low Income Economies. It is remarkable that the first pilot project of the GO::DH, "AroundDH in 80 Days" could immediately fill the list of DH projects around the world (Gil) with the help of international volunteers. In the context of the humanities, globalization is not always intrinsically good, but international communication would be significant for DH and the humanities.

While there have been efforts focused actual local development, in some cases, such as of the Japan, most DH activities hadn't been known in the global community and most of global DH activities hadn't been known in Japan until several years ago. This is in spite of the fact that the number of identified Japanese DH-related researchers is over 200 and recently the domestic annual DH conferences have gathered 40-60 papers every year, with 800 papers being presented since 1989 from many universities, museums, libraries, and other institutions in a DH-related quarterly workshop (A. Charles Muller). As the case might be similar in other non-English and non-Western countries, it might be useful to report our recent attempts to bridge between the DH research being carried out by non-English-speaking scholars and those in the international community.

First, the establishment of the JADH has proved itself to be one of the most effective solutions for closing this kind of gap. Since around 80 researchers participated in the first conference in 2011 in Osaka, 80-90 Japanese researchers have attended the annual conference and communicated with international researchers. Then, several germs of international collaboration have come into being there and Japanese researchers who paid attention to the results of research activities of the international DH community have gradually increased as a "methodological commons"—although most of the research is still focused on Japanese or Eastern materials.

Secondly, an e-newsletter titled "Digital Humanities Monthly" (DHII and ARG) has been published by the International Institute for Digital Humanities since July 2011. It has 390 subscribers and is also published on the Web. The e-newsletter written in Japanese consists of an invited essay, brief news of international activities of DH and Digital History, DH-event calendar, and reports of DH events held in Japan and foreign countries collaborating with some local and international voluntary DH researchers. The event reports are plotted on a time-space map of the Neatline.(fig.1) The total number of the access to the Web pages was over 5,000 this October. According to comments of the readers, it seems to have gained the attention of not only DH researchers, but also librarians, curators, archivists, publishers, and general public, enabling them to see the picture of the entire situation of the domestic and the international DH.

Thirdly, we plan to make it easier to treat Japanese and Eastern materials compliant with kinds of international standards. So far, we are working to propose the encoding of Han characters that occur in our research materials in the Universal Character Set as a group of researchers (rather than as a national body, as has been the policy heretofore) so that researchers can not only treat the characters but also propose the inclusion of new characters more easily. Moreover, we are planning to form an appropriate guideline of text encoding of Japanese and Eastern materials in the framework of the Text Encoding Initiative P5 guidelines (Bauman) collaborating with related researchers around the world. As a preparation for this, we've held full-day TEI workshop by the participation of international TEI researchers over 10 times and taught the framework of the TEI to 50 researchers in total.

While it has up to now been difficult to bridge the local and the global, we hope our attempts will be useful for an

appropriate mode of globalization. We would like to discuss various possibilities with participants in the conference.



Fig. 1:

References

- A. Charles Muller, Kozaburo Hachimura, Shoichiro Hara, Toshinobu Ogiso, Mitsu Aida, Koichi Yasuoka, Ryo Akama, Masahiro Shimoda, Tomoji Tabata, and Kiyonori Nagasaki (2010). "The Origins and Current State of Digitization of Humanities in Japan." Digital Humanities 2010 : 68-70.

Bauman, Lou Burnard and Syd (2007). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html>.

DHII and ARG. Ed. Kiyonori Nagasaki et al. 7 (2011). www.dhii.jp/DHM .

Gil, Alex ed. AroundDH | Global List. (2013). <http://goo.gl/4ov5nR>.

Active Authentication through Psychometrics

Noecker Jr, John

Juola & Associates, United States of America

What can your computer habits reveal about you? The answer might surprise you. Previous work (Juola, et al., 2013) has shown that just a few minutes of computer usage can be used to identify who is at the keyboard and their demographic and psychological attributes with a fairly high degree of accuracy. We expand upon this to show that the same usage data can be used to thoroughly profile a previously-unknown user to obtain valuable psychological information about the user.

Authorship attribution, the analysis of a document's writing style to infer the author's identity, is a well-established problem in text classification. Previously, we used classical authorship attribution techniques to identify "who was at the keyboard" using the DARPA Active Authentication Corpus (Juola et al., 2013). Researchers have successfully applied the analysis of language usage to infer authorship of written documents (Juola, 2006. Koppel et al, 2009. Stamatatos, 2009. Jockers & Witten, 2010), and stylometric analysis has also been applied to things like gender (Argamon et al, 2006), personality (Luyckx & Daelemans, 2008), and even psychological disorders like depression (Rude et al. 2004).

Here, we attempt to perform the same technique with groups composed of individuals who share common psychological traits. Previous work (Luyckx & Daelemans, 2008; Noecker & Juola, 2013) on personality profiling has so far focused on analyzing previously-written documents. In contrast, our system provides a method for real-time psychological profiling of a user based on his or her interactions with a computer over a relatively short period of time (approximately 30 minutes).

The ultimate goal is two-fold: to learn something about a previously-unobserved user (traditional stylometric identification techniques require us to have training data on a user before

we can identify him) and to use psychological traits as an enhancement to current user authentication methods.

Currently, exact accuracy on the user-based authentication is approximately 90%. This task becomes more difficult (and the accuracy becomes correspondingly lower) as the pool of potential author models grows. In order to improve overall accuracy of the user authentication task, we propose to include these psychological profiling tools in the authentication system. If a given user can be identified as the most likely candidate with 90% probability, and several facets of that user's personality can be confirmed with similarly high confidence, this will increase the overall robustness of the authentication system.

For our purposes, we used two personality/intelligence measurement systems to profile users: Myers-Briggs Type Indicator (MBTI) and Multiple Intelligences Developmental Assessment Scales (MIDAS).

The Myers-Briggs type indicator (MBTI) assigns four binary classifications to define personality (Myers & Myers, 1980)

- Extroversion vs Introversion
 - iNtuition vs Sensing
 - Thinking vs Feeling
 - Judgement vs Perception

The Multiple Intelligences Developmental Assessment Scales (MIDAS) were developed by Dr. Howard Gardner in his 1983 book "Frames of Mind" (Gardner, 1983). He used a unique definition of intelligence: "The ability to solve a problem or create a product that is valued within one or more cultures" (MI Research and Consulting). He identified 8 primary intelligent scales, each of which have several subscales (MI Research and Consulting):

- Musical
 - Vocal Ability
 - Instrumental Skill
 - Composer
 - Appreciation
 - Kinesthetic
 - Athletics
 - Dexterity
 - Logical-Mathematical
 - Everyday Math
 - School Math
 - Everyday Problem Solving
 - Strategy Games
 - Spatial
 - Space Awareness
 - Working with Objects
 - Artistic Design
 - Linguistic
 - Expressive Sensitivity
 - Rhetorical Skill
 - Written-academic
 - Interpersonal
 - Social Sensitivity
 - Social Persuasion
 - Interpersonal Work
 - Intrapersonal
 - Personal Knowledge / Efficiency
 - Effectiveness
 - Calculations
 - Spatial Problem Solving
 - Naturalist
 - Animal Care
 - Plant Care

We also include a 9th main scale, *Leadership*, with its own subscales: *Communication*, *Management*, and *Social*.

Materials and Methods

Corpus

In order to create the most accurate corpus possible, we set up a simulated office environment and hired 80 temporary workers for one week each. Workers were tasked to perform a long-term blogging project (research and write blog articles on topics “related to Pittsburgh in some way”) over the course of a normal workweek. For this study, we use the Free Key Logger output, which provides the exact text typed by each user. We do not include any information about the applications being used or any data the user pastes from the clipboard.

Feature Extraction

For our analysis, we used the Java Graphical Authorship Attribution Program (JGAAP) (Juola et al, 2009). JGAAP is a Java-based, modular program for textual analysis, text categorization, and authorship attribution. It provides a comprehensive framework, allowing us to rapidly test the effectiveness of different analysis techniques on the recorded data.

JGAAP divides analysis into several steps: Canonicalization (Preprocessing), Event Set (Feature) Generation, and Analysis. In Canonicalization, preprocessors are used to standardize the text. For this step, we converted all input letters to lower case (“Unify Case”) and converted all strings of whitespace characters into a single space character (“Normalize Whitespace”). At this stage, we also processed a variety of special keyboard characters, converting these non-printable characters into a printable placeholder (e.g. “backspace” was replaced with “B”). Finally, we divided the input data into blocks of 1,000 characters, representing about 30 minutes of computer usage.

For the event set generation, we tested character N-grams for all N from 1 to 15, and word N-grams for N from 1 to 5.

We then applied a number of analysis methods for each experiment: Cosine Distance, Intersection Distance, Manhattan Distance, and Matusita Distance. For each method, we used a centroid-based nearest neighbor classifier. We performed leave-one-out cross-validation to reach our final conclusion.

Models

For the MBTI classifiers, we built four binary classifiers (i.e. E vs I, N vs S, T vs F, and J vs P). For the MIDAS classifiers, we first built a single 9-way classifier to identify a user’s principle main scale. This was the scale along which the user scored highest (i.e. the scale for which the user showed the highest preference). For example, a user might have a preference for “Musical” or “Linguistic”. We also developed subscale classifiers, which identify a user’s preference within each major scale. For instance, a user might be identified as “(Musical) Vocal Ability” and “(Kinesthetic) Dexterity”, etc. Thus, each user was identified by a single main scale preference as well as nine subscale preferences.

Results

MBTI

For the MBTI classifiers, we averaged an accuracy of 81.5%. The expected baseline average (assuming we pick the most prevalent personality type for each category) is 55%.

MIDAS

For the MIDAS main category identification, our best performing classifier had accuracy of 70.7%. This was using character 15 grams with Intersection Distance. The expected baseline accuracy (achieved by choosing the most common main scale, “Linguistic”) was 22.1%.

For the MIDAS subscale identification, the best performing classifiers used a variety of Character n-grams, again with Intersection Distance as the top performing analysis method. The average subscale accuracy was 81%.

Conclusion

We have shown here a method to reliably psychologically profile a computer user based on only a short period (about 30 minutes) of usage time. In addition to providing valuable information about the user in question, this method can also be used to provide additional layers of security for the active authentication system we have described previously. Even in an adversarial situation, the difficulty of imitating both an individual user’s style, as well as mimicking the psychological profile of the user, will provide additional security to the authentication system.

Also interesting to note is the limited usage data required to perform these analyses. The initial user psychological testing period took approximately 3 hours, but accurate results were obtained for only 30 minutes of computer usage. In addition, the three hours of testing were completely lost time – the users were able to work only on the tests during this time. In contrast, the 30 minutes of analysis can be done on whatever the user is working on at the time. No downtime is required to perform these analyses. We believe this system could be useful anywhere a non-intrusive analysis of a user might be beneficial (e.g. determining whether a potential employee would be a good fit).

For future work, we intend to focus on reducing the amount of data needed even further. Preliminary results on as little as 500 characters (about 15 minutes of usage time) have been promising. Additional work is also being done to integrate these methods into the broader active authentication system in order to bolster the overall reliability of the system.

References

- Argamon, S., Koppel, M., Fine, J., Shimoni, A. R (2006).** “*Gender, Genre, and Writing Style in Formal Written Texts*”. Interdisciplinary Journal for the Study of Discourse. Volume 23. Issue 3. pp. 321-346.
- “*Broad Agency Announcement: Active Authentication*”. (2012). DARPA. *Solicitation No. DARPA-BAA-12-06*. 12 Jan. 2012. <http://www.fbo.gov/index?tab=documents&t_abmode=form&subtab=core&tabid=494b6b2c612c4fd3db6cb018d4467e21>.
- Gardner, Howard (1983).** “*Frames of Mind: The Theory of Multiple Intelligences*”. Basic Books.
- Jockers, M. L., Witten, D. (2010)** “*A Comparative Study of Machine Learning Methods for Authorship Attribution*”. Literary and Linguistic Computing. vol. 25, no. 2. pp. 215–23.
- Juola, P. (2006)** “*Authorship Attribution*”. Foundations and Trends in Information Retrieval, vol. 1, no. 3. pp. 233–334.
- Juola, Patrick, Noecker Jr., John, Ryan, Mike, Speer, Sandy (2009).** “*JGAAP 4.0 – A Revised Authorship Attribution Tool*”. Proc. Digital Humanities 2009. pp. 357–359. Maryland Institute for Technology in the Humanities. University of Maryland.
- Juola, Patrick, Noecker Jr., John, Stolerman, Ariel, Ryan, Michael, Brennan, Patrick, Greenstadt, Rachel (2013).** “*Keyboard Behavior Based Authentication for Security*”. IT Professional. 18 June 2013. IEEE computer Society Digital Library. IEEE Computer Society. <<http://doi.ieeecomputersociety.org/10.1109/MITP.2013.49>>.
- Koppel, M., Schler, J., Argamon, S. (2009)** “*Computational Methods in Authorship Attribution*”. J. Amer. Soc. Information Science and Technology, vol. 60, no. 1. pp. 9–26.
- Luyckx, K., Daelemans, W. (2008)** “*Personae, a Corpus for Author and Personality Prediction from Text*”. Proceedings of the Sixth International Conference on Language Resources and Evaluation. Marrakech, Morocco.
- MI Research and Consulting, Inc.** “*Multiple Intelligences Theory*”. <www.miresearch.org/mi_theory.html>.

- Myers I B, Myers P.** (1980) "Gifts Differing: Understanding Personality Type". Palo Alto, CA. Consulting Psychologists Press.
- Noecker Jr., J. Juola, P.** (2013) "Psychological Profiling Through Textual Analysis". Literary and Linguist Computing.
- Rude, S., Gortner, E., Pennebaker, J.** (2004) "Language Use of Depressed and Depression-Vulnerable College Students". Cognition and Emotion.
- Stamatatos, E.** (2009) "A Survey of Modern Authorship Attribution Methods". J. Amer. Soc. Information Science and Technology, vol. 60, no. 3. pp. 538–556.
- Zheng, N., Paloski, A., Wang, H.** (2011) "An Efficient User Verification System via Mouse Movements," Proc. 18th ACM Conf. Computer and Communications Security (CCS 11). ACM. pp. 139–150.

Encoding Metaknowledge for Historical Databases

Nuessli, Marc-Antoine

nuessli.ma@gmail.com

EPFL

Kaplan, Frédéric

frederic.kaplan@epfl.ch

EPFL

Motivation

Historical knowledge is fundamentally uncertain. A given account of an historical event is typically based on a series of sources and on sequences of interpretation and reasoning based on these sources. Generally, the product of this historical research takes the form of a synthesis, like a narrative or a map, but does not give a precise account of the intellectual process that led to this result.

Our project consists of developing a methodology, based on semantic web technologies, to encode historical knowledge, while documenting, in detail, the intellectual sequences linking the historical sources with a given encoding, also known as *paradata*¹. More generally, the aim of this methodology is to build systems capable of representing multiple historical realities, as they are used to document the underlying processes in the construction of possible knowledge spaces.

Overview of the Approach

Semantic web technologies, with formal languages like *RDF* and *OWL*, offer relevant solutions for deploying sustainable, large-scale and collaborative historical databases (see for instance²). Compared to traditional relational databases, these technologies offer more flexibility and scalability, avoiding the painful problems of large schema migration. They are grounded in logic and thus permit us to easily conduct semantic inferences. Some very stable semantic based ontologies like *CIDOC-CRM*, now an ISO standard, have been used successfully in the cultural heritage domain for about 20 years³.

However, the languages used in the semantic web technologies have a major limitation that prevents their usage for encoding metahistorical information. Expressed knowledge is typically formalised with *RDF* triplets which are not objects in the same order as the knowledge content (*RDF* resources identified with *URIs*) to which they link. For example, it is difficult to document the source, the author or the uncertainty of given *RDF* statement.

One way to compensate for this flaw, while respecting the W3C norms, consists of transforming each *RDF* triplet (*subject predicate object*) into three triplets (*statement rdf:subject subject*), (*statement rdf:predicate predicate*), (*statement rdf:object object*). Using this approach, it becomes

possible to add new triplets with a given statement as subject, documenting additional paradata about this statement. The resulting knowledge base can include metahistorical information, i.e. information about historical information creation processes. This metainformation can document the choice of sources, transcription phases, coding strategies, interpretation methods and whether these steps are realised by humans or machines. Thus, each historical database designed following this methodology integrates two levels of knowledge. The first level provides the documentation about the origin, the nature and the formalisation used to encode historical data, while the second level codes for the historical data itself.

The Knowledge Construction Vocabulary (KCV)

We are working on a specific *RDF* vocabulary, called *Knowledge Construction Vocabulary* (KCV), which will enable us to implement the two level organisation using the standards of the semantic web. KCV *RDF* statements represent knowledge construction steps, while effective historical knowledge is only expressed through reified triplets. An important concept in this vocabulary is the notion of *knowledge spaces*. A knowledge space designates a closed set of coherent knowledge, typically based on a defined set of sources and methods. Examples of knowledge spaces include documentary spaces (e.g. a defined corpus of sources) and fictional spaces (e.g. a coherent world typically described in a book).

Figure 1 shows an example of the kind of graphs that can be built using the KCV vocabulary. In this example, two knowledge spaces have been defined: one documentary space (*DHLABDocuments*) and one so-called fictional space (*HistoireVenise_S1*). Each of these two spaces is defined as a unique resource with an associated URI. A statement (*Statement1*) stands for a reified triplet defining that (*HistoireVenise*) is a kind of *Book* and is linked to the documentary space. The KCV vocabulary allows us to document who entered the information (*fournier*) and the creation time of the statement (*May 06th*). To formalise the fact that the book, *HistoireVenise*, is used as a knowledge source, a specific resources *HistoireVenise_KS* is created and linked with the *HistoireVenise*, the book, and the general document space *DHLABDocuments*.

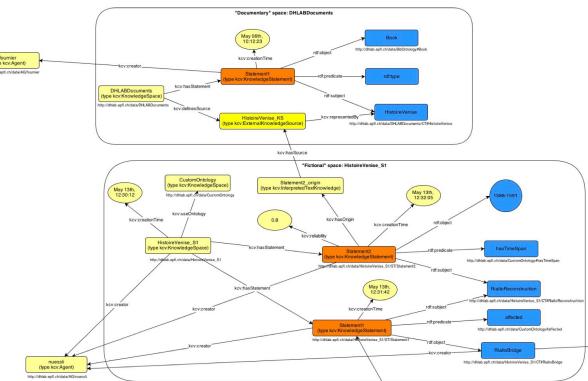


Fig. 1: A "toy" example of the use of the KCV vocabulary to code historical and metahistorical information

In the fictional space *HistoireVenise_S1*, a statement (*Statement2*) codes for a reified triplet indicating that the reconstruction of the Rialto bridge occurred during the period of 1588-1591. Information about the author, the creation date and the reliability of *Statement2* are documented using various KCV triplets. The link between the document space and the fictional space is encoded by a link between the knowledge source *HistoireVenise_KS* and a statement, *Statement2_origin*, linked to *Statement2* of type *interpretedtextknowledge*.

We can make three remarks:

1. This is obviously a "toy" example (real graphs encoding historical data are typically much bigger), but it illustrates how historical and metahistorical information can be coded

with a linked data approach. This allows us to envision queries mixing both historical and metahistorical requests, for instance reconstructing an historical context based only on certain kinds of sources or excluding information that was provided to the database by some authors.

2. The kind of intellectual processes documented by *KCV* can easily include algorithmic steps like digitisation, optical character recognition pipelines on documents, text mining, semantic disambiguation, etc. The version and the author of the algorithms used can easily be included using *KCV* statements. This kind of documentation permits us to exclude historical information linked with some processing using early versions of algorithms that may have "polluted" the data. This is an important prerequisite for building sustainable databases in the long term.
3. Documenting metahistorical information using *KCV* may look like a tedious process; however, in most cases, this information can be inserted automatically using a higher-level interface. A database interface in which the user is logged permits to easily produce historical data based on the *KCV* vocabulary, taking the form of reified *RDF* triplets, while documenting the author, the data and the methods used.

Ontologies Matching

The *KCV* approach for encoding historical databases is also interesting from the perspective of ontologies alignment: a notoriously difficult issue⁴. Each research group tended to code historical data using their own local ontologies, adapted to their research approach. The metahistorical documentation provided by the *KCV* vocabulary enables us to envision strategies for mapping such ontologies to a pivot ontology. Figure 2 shows this general process in which several knowledge spaces are linked. Each group locally describes the source documents used (1), transcribes their content (2) and eventually codes/interprets this content (3). Throughout this process, two groups produced two independent custom ontologies (A and B). The alignment process proceed in two additional steps. First, both local ontologies are mapped onto a general content ontology (4) (for instance *CIDOC-CRM*, but not necessarily) and then, once expressed in this common conceptual model, the information contained in the graph is aligned and the content is merged (5).

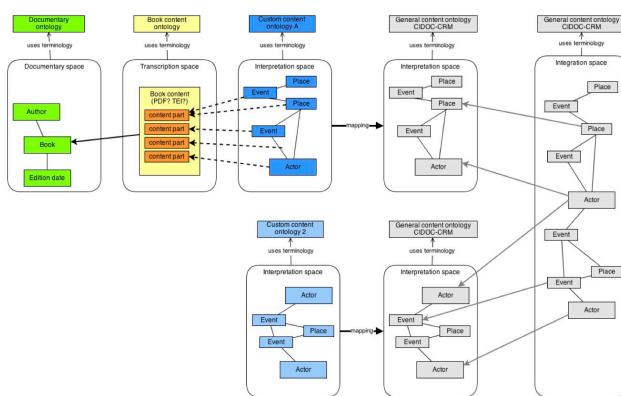


Fig. 2: The general process of ontologies matching

Figure 3 gives a more detailed account of the final step. First knowledge sources are mapped, then types are mapped and eventually predicates are mapped. In some cases, only a partial level of correspondence can be reached. These steps can be done manually or automatically and are, of course, subject to errors. It is therefore crucial to document the authors of these matching steps, whether they are humans or algorithms. This is why the authors are, linked all the other steps, described in the *KCV* vocabulary.

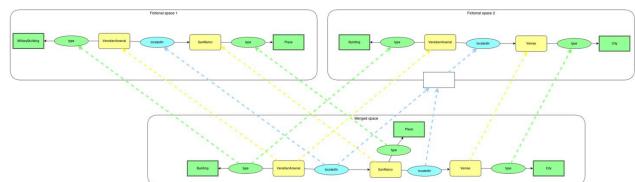


Fig. 3: Detail of the ontologies matching process

Conclusion

The approach briefly presented in this paper enables us to encode historical and metahistorical data in a unified framework. The method we describe is fully compliant with the current technologies and standards of the semantic web (*RDF*, *SPARQL*, etc.). It does impose a unified historical terminology but can also be used in conjunction with existing standards. For instance *CIDOC-CRM* can be used to describe historical knowledge extracted from archival documents (e.g events, people, places) using *RDF* triplets and *KCV* can be used to code information about the *CIDOC-CRM* triplets themselves, such as documenting who entered a particular triplet. The originality of our proposal comes from the introduction of the second level (metahistorical) on top of the existing *RDF* ontologies. This does not necessarily impose an additional burden on the person encoding the historical data. Using a dedicated web interface, the metahistorical information can be added automatically as the data is progressively entered.

Coding metahistorical information by making explicit the many underlying modelling processes allows us to prepare for possible ontology evolution and enables easier ontology matching. More importantly, our approach does not impose the search for a global truth (a unique and common version of historical events) but pushes towards the explication of the intellectual and technical processes involved in historical research, thus giving the possibility of fully documented historical reconstructions.

References

1. Bentkowska-Kafel A., Denard H., and Baker D. (2012). *Paradata and Transparency in Virtual Heritage*. Ashgate Publishing, Ltd.
2. Ide, N., and D. Woolner (2007). *Historical Ontologies*. Words and Intelligence II: 137–152.
3. Doerr, M. (2003) *The CIDOC CRM – an Ontological Approach to Semantic Interoperability of Metadata*. AI Magazine 24, no. 3.
4. Shvaiko P. and J. Euzenat (2013). *Ontology matching: state of the art and future challenges*. IEEE Transactions on Knowledge and Data Engineering, 25(1): 158-176.

Two Irish Birds: A Stylometric Analysis of James Joyce and Flann O'Brien

O'Sullivan, James

josullivan.c@gmail.com
University College Cork, Ireland

Bazarnik, Katarzyna

Jagiellonian University, Krakow

Eder, Maciej

Pedagogical University of Krakow

Rybicki, Jan

Jagiellonian University, Krakow

It has long been argued that Brian O'Nolan, operating under the pseudonym of Flann O'Brien, is a disciple of James Joyce. This paper examines the stylometric similarities between the

two authors, particularly in relation to *At SwimTwoBirds* and, to a lesser extent, *The Hard Life*, which we demonstrate are stylistically the most Joycean novels from O'Brien's oeuvre. Emerging from a wider analysis of modernist writers (O'Sullivan, 2014), we will outline the results of a series of quantitative enquiries focused specifically on Joyce and O'Brien, before offering a number of literary interpretations.

O'Brien's *At SwimTwoBirds*, despite considerable critical acclaim, was initially illreceived as a product of its "Joycean undertones", commentators "tend[ing] to condemn the work as inferior imitation" (Hopper, 1995: 46). Séán Ó Faoláin remarked that the novel had "a general odour of spilt Joyce all over it", while the *New Statesman* branded it as "dull" on account of its "long passages in imitation of the Joycean parody" (*ibid.*). Asbee, while critical of the *The Hard Life*, accepts that some comparisons can be drawn between it and Joyce's collection of short stories, *Dubliners* (Asbee, 2001).

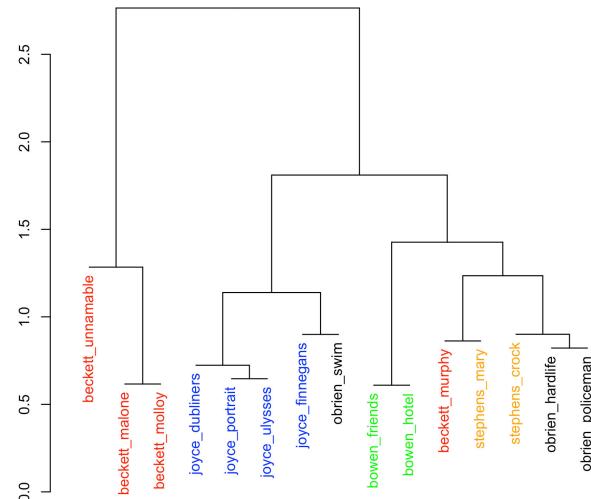
The relationship between O'Brien and Joyce remains a concern for scholars. Hopper argues that "O'Brien is usually lumped in with Joyce" as a result of "their historical and cultural proximity", but that this is "an assumption which is unfair to both writers" (Hopper, 1995: 14). Stylistically, O'Brien's novels are littered with parodic tributes to Joyce (O'Grady, 1989). Indeed, while O'Brien demonstrated "repeated efforts to escape his influence" (Dotterer, 2004: 59), "*At Swim* had everything in the world to do with James Joyce" (Taaffe, 2004: 253). Some critics maintain that the "omnipresence of Joyce ... was to be expected" on account of O'Nolan's shared affiliation with University College Dublin (*ibid.*: 249). While Joyce may have been a "talismanic figure" at UCD (*ibid.*: 249), O'Brien's Joycean parodies are not always interpreted as positive. Taaffe suggests that O'Brien's "attitude towards the elder writer ... is equivocal, at the very least" (*ibid.*: 253), while McMullen argues that "*At SwimTwoBirds* enters into dialogue not with James Joyce alone" (McMullen, 1993: 63). Dotterer aptly summarises this debate: "Critical comparison with Joyce has been frequent, as have analytical comparisons of their fiction, but less often has an awareness of this link to Joyce been seen as central and persistent in Brian O'Nolan's formation of his own work. This link with James Joyce was one O'Nolan embraced, at times begrudgingly or unwillingly, but always out of some inner artistic and psychic necessity" (Dotterer, 2004: 54).

By offering a fresh appraisal based on quantitative methods, this paper identifies the specific points at which O'Brien's Joycean parodies are most prominent, so that literary interpretations can be focused, with computational precision, on the relevant passages.

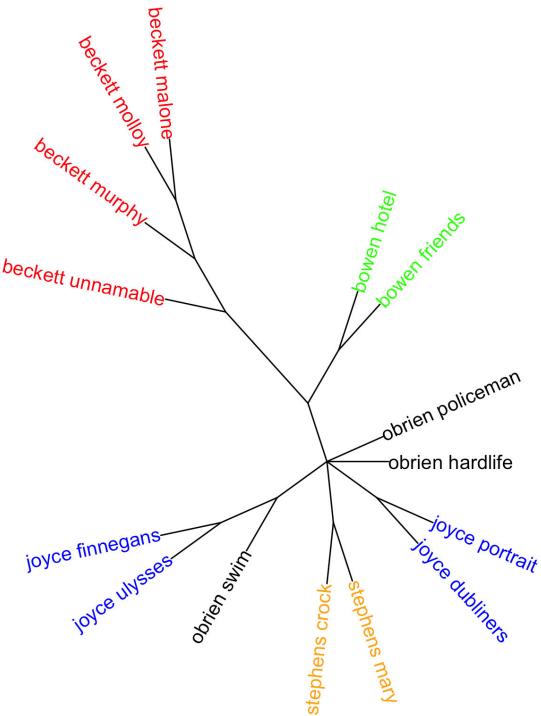
Methodology & Results

A number of multivariate stylometric methods were used in this study. Cluster Analysis provided a preliminary insight into the dataset, identifying main groupings. Since Cluster Analysis is very sensitive to the number of features (most frequent words) analysed, the next step involved generating Bootstrap Consensus Trees, or dendograms averaging numerous single Cluster Analysis trees. We measured the 100 most frequent words, expanding this range from 100 to 1000 in intervals of 100 in order to produce a number of virtual dendograms combined into one consensus plot. The distance measure in all the tests was derived from Burrows's Delta (Burrows, 2002; Hoover, 2004). Finally, to identify (possible) peculiarities in sequential development of the analyzed texts, we used Rolling Delta (Rybicki et al., 2013), which forms an authorial signature based on one set of texts, and then applies that fingerprint to another text. Authorial signatures are then plotted over the text in question, with stylistic similarity indicated through proximity to the baseline. The aforementioned methods were applied using the R package "stylo" (Eder et al., 2013).

Initially, a cluster analysis was generated using a selection of English language Irish modernists. Using the 100 most frequent words, with 100% culling, it was interesting that O'Brien's *At SwimTwoBirds* clustered with Joyce's texts:

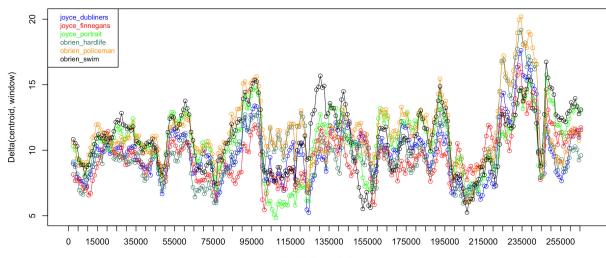
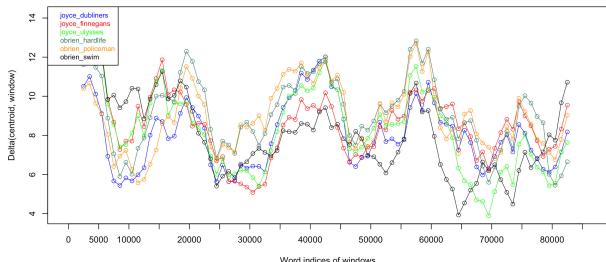
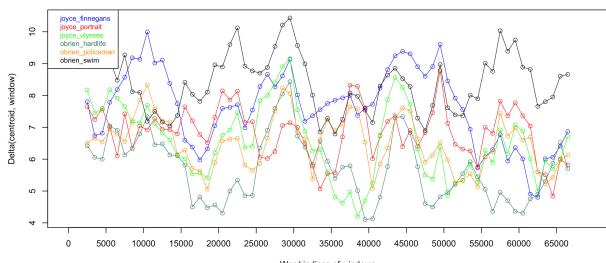


This prompted further exploration, so the Bootstrap Consensus Tree, a more robust measure of style, was conducted. As can be seen, O'Brien's novels continued to cluster with Joyce:



With our cluster analysis and bootstrapping confirming the common belief that O'Brien's style was strongly influenced by Joyce, we adopted Rolling Delta (Rybicki et al., 2013) as a means of pinpointing specific passages of interest within the relevant corpora. The most significant findings are as follows:

Rolling Delta analysis, *Ulysses*:

Rolling Delta analysis, *A Portrait of the Artist as a Young Man*Rolling Delta analysis, *Dubliners*:

As evidenced above, there are a number of places in these texts where O'Brien's authorial signature is particularly clear. Thus, we can identify these sections as areas of a distinct crossover between the style of the two authors. O'Brien's idiom of *At SwimTwoBirds* emerges, quite strongly, in two sections of *Ulysses*, and in several sections of *A Portrait of the Artist as a Young Man*. Interestingly, *The Hard Life* is stylistically similar to *Dubliners* throughout, consistently more so than any of Joyce's other texts. Specific locations within the texts were identified using the following command in BASH:

```
awk '{for(i=N;i<M;i++) print $i}' RS= ulysses.txt
which prints everything between the Nth and Mth word in the file.
```

Literary Interpretations

These results contribute significantly to scholarship surrounding Joyce and O'Brien in that they offer a clear picture of where the style of both authors are most similar. Below, we will give specific focus to correlations between *At SwimTwoBirds*, *Ulysses* and *A Portrait*, as well as *The Hard Life* and *Dubliners*.

At SwimTwoBirds

Our Rolling Delta analysis demonstrates significant similarities between the style of *At SwimTwoBirds* and the "Oxen of the Sun" and "Eumeaus" episodes in *Ulysses*. Interestingly, "Oxen" and "Eumeaus" are stylistically distinct

in that they both offer parodies based on language: in "Oxen" the parody is centred around various literary figures, in "Eumeaus", the focus is on the bourgeois. Incidentally, "Oxen" and "Eumeaus" are among the few episodes in which Stephen and Bloom appear together. Thus, two interpretations present themselves: firstly, that the results of our analysis can be attributed to O'Brien's imitation of the Joycean parody, of which "Oxen" and "Eumeaus" are archetypal. Joyce's exaggerated style in "Oxen" parodies the chronological progression of the English literary canon from Early English to Twentieth Century slang. Very much a Menippean satire, *At Swim* is intensely parodic, and like "Oxen", draws upon a wide range of sources from "high" modernist works to correspondence with a horse racing pundit.

Alternatively, the presence of Stephen and Bloom may be accountable for the results, the product of their distinct correlation with the *At Swim* characters. "Oxen of the Sun" is the first episode in which Stephen and Bloom appear together, while they are also both present in "Eumeaus". O'Brien's unnamed protagonist in *At Swim* has long been considered a revival of Joyce's artist, personified in the figure of Stephen Dedalus, hence possibly Stephen's presence in these passages is a key. However, in both episodes, Bloom's consciousness seems more prominent, while the earlier episodes, where Stephen features more heavily, show little proximity to O'Brien's style. We could conclude from this that connections between the young artists in *At Swim* and *Ulysses* are more symbolic than stylistic. An exception to this finding potentially exists in the *Portrait*, where the style of *At SwimTwoBirds* is very similar to the final sections of Joyce's first novel, which are dominated by a maturing Stephen who appears more assured in his positions and moral development. Much has been said on the nature of Stephen's progression from *A Portrait* to *Ulysses*; our findings would suggest that O'Brien's student has more in common with the Stephen who is looking to "fly by those nets" (Joyce: 231) than with the Stephen we encounter in Joyce's epic. A triad of stylometric connections emerges at this juncture. Firstly, in "Eumeaus" Joyce presents a parody of bourgeois attempts at sounding cultured. Besides, a stylistic similarity may be connected with the ironic distance with which Joyce writes Stephen in later parts of the *Portrait* and "Eumeaus". Thus, from the perspective of style, we can conclude that O'Brien offers a similar treatment of the bourgeoisie in *At SwimTwoBirds*. Another interpretive possibility is connected with W.B. Murphy and SkintheGoat Fitzharris, two storytellers in "Eumeaus", weaving fantastic tales in rambling style, which may find parallels in *At Swim*. In fact, stylistic similarities between these Ulyssean episodes and O'Brien's novel may be due to their polyphonic (in the Bakhtinian sense) texture rather than affinities between styles of particular heroes. It is hoped that a more detailed stylometric positioning of similar passages, combined with their close reading, will verify the above hypotheses.

The Hard Life

O'Brien's tendency to present an archetypal Dublin dialect across many of his novels is another possible explanation for his close proximity to Joyce's style. Clune argues that it was O'Brien's Ulster Irish that "sharpened his ear for Dublin dialect" and let him "capture the precise nuances of Dublin speech. He himself claimed that Joyce had the edge on him in this, but there are those who disagree, who argue that only a non-Dubliner could have 'caught' his Dubliners so precisely, pinning them down 'phrase by phrase' as he put it himself" (Clune, 1986: 6). Indeed, "Dublin dialogue has a special relish for Brian O'Nolan", and he praised Joyce for the "supernatural skill" which he wrote such (Mays, 1974: 246). Thus, it is perhaps unsurprising that both writers' affection for Dublin dialect results in their styles being so similar.

While most of O'Brien novels were centred around Dublin, it is *The Hard Life* which is closest to *Dubliners*. Published 47 years after Joyce's collection, the proximity of O'Brien's style to that of *Dubliners* demonstrates that O'Brien, though not a Dubliner himself, mastered a style long dominated by Joyce. This is counter to much of the novel's criticism, which

accuses O'Brien of being the overt protégé, too conscious in his attempts at achieving the ideal Joycean parody. Asbee suggests that comparisons between *The Hard Life* and Joyce's work is "almost insulting" (Asbee, 2001). While O'Brien may be charged with repeated imitation of Joyce, our analysis illustrates, and this paper will discuss in the context of stylometry, why, in some instances, he cannot be dismissed as having failed in his attempts.

References

- Asbee, S.** (2001). *Flann O'Brien*. Boston: Twayne Publishers.
- Burrows, J.** (2002). 'Delta': a measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17(3): 267–87.
- Clune, A.** (1986). *Flann O'Brien: twenty years on*. The Linen Hall Review, 3(2): 4–7.
- Dotterer, R. L.** (2004). *Flann O'Brien, James Joyce, and 'The Dalkey Archive'*. *New Hibernia Review / Iris Eireannach Nua*, 8(2): 54–63.
- Eder, M., Kestemont, M. and Rybicki, J.** (2013). *Sylometry with R: a suite of tools*. In: Digital Humanities 2013: Conference Abstracts. Lincoln (NE): University of NebraskaLincoln, pp. 487–89.
- Hoover, D. L.** (2004). Testing Burrows's delta. *Literary and Linguistic Computing*, 19(4): 453–75.
- Hopper, K.** (1995). *Flann O'Brien: A Portrait of the Artist as a Young Postmodernist*. Cork: Cork University Press.
- Joyce, J.** (1996). *A Portrait of the Artist as a Young Man*. London: Penguin Books.
- Mays, J. C. C.** (1974). *Brian O'Nolan and Joyce on art and on life*. *James Joyce Quarterly*, 11(3): 238–56.
- McMullen, K.** (1993). Culture as colloquy: Flann O'Brien's postmodern dialogue with Irish tradition. *Novel: A Forum on Fiction* 27(1): 62–84.
- O'Grady, T. B.** (1989). High Anxiety: Flann O'Brien's *Portrait of the Artist*. *Studies in the Novel*, 21(2): 200208.
- O'Sullivan, J.** (2014). Modernist frequencies: A computational stylistics approach to national Modernisms. In: *MLA 2014: Conference Proceedings*. Chicago, forthcoming.
- Rybicki, J., Kestemont, M. and Hoover, D. L.** (2013). Collaborative authorship: Conrad, Ford and rolling delta. In: Digital Humanities 2013: Conference Abstracts. Lincoln (NE): University of NebraskaLincoln, pp. 368–71.
- Taaffe, C.** (2004). 'Tell me this, do you ever open a book at all?': portraits of the reader in Brian O'Nolan's 'At SwimTwo-Birds'. *Irish University Review*, 34(2): 247–60.

Modeling Linguistic Research Data for a Repository for Historical Corpora

Odebrecht, Carolin

Humboldt-Universität zu Berlin

Existing historical linguistic corpora vary a great deal with respect to formats, corpus architecture, annotation types and values and preparation steps. The LAUDATIO-Repository (www.laudatio-repository.org) provides an open access environment to facilitate the management of such heterogeneous research data with an extensive, uniform and structured documentation and faceted and free-text search without limitation with respect to formats or annotations. For this purpose, we have developed a meta-model which is expressive enough to represent a large variety of corpus formats. This meta-model, described as a TEI-ODD specification with automatically generated schemas (Burnard & Rahtz 2004), is also the basis for the technical implementation in the repository.

Building and analyzing historical corpora often incorporates diplomatic transcriptions, normalizations of these transcriptions and research specific annotation layers which will be illustrated with the help of two corpora; the German Manchester Corpus

(GerManC) and the RIDGES Herbology Corpus. GerManC¹ (Durrell, Ensslin & Bennett 2007) contains for instance two formats (TEI XML and CoNLL) which represent different kinds of annotations and analyses. The TEI XML format contains a diplomatic transcription and a register specific mark-up. By contrast, the CoNLL format contains token annotations for normalization, part-of-speech (POS) and lemmatization (e.g. the STTS tag set, Schiller et al. 1999) as well as morphology and dependency annotation for syntactic relations between the tokens (e.g. Foth 2006). Thus, this corpus uses two formats for encoding different kinds of annotations and analyses.² On the other hand, the second version of the RIDGES Herbology Corpus³ contains all annotations in one format (EXMARaLDA, Schmidt & Wörner 2009) which is then converted into the relANNIS format used by the ANNIS corpus system (Zeldes et al. 2009) for search and visualization capacities. The corpus architecture of RIDGES is specific in the following way: Via multiple segmentations, annotations can refer to different basic textual data in the corpus (Krause et al. 2012). To normalize separate spellings of complex verbs in historical German such as *zusammen gesetzt* to *zusammengesetzt* (RIDGES, Curioser Botanicus oder sonderbares Kräuterbuch, 1675), the tokens need to be merged in the normalized annotation whereas tokens need to be separated when normalizing *zuverstehen* to *zu verstehen* (RIDGES, Alchemistica Praktik 1603). Every further annotation — for instance the POS annotation may either refer to the diplomatic segmentation layer or to the normalized segmentation layer.

Having identified what exactly needs to be described by a meta-model, we then define the actual use-cases associated with this meta-model. With respect to range, specificity and user scenarios, distinct requirements could only be designed for concrete applications. For this study, the LAUDATIO-Repository (Krause et al. 2013) is taken as an example. In this case, the meta-model will enable a retrieval of, a structured search on and a holistic and extensive documentation of the heterogeneous historical corpora and their preparation (for further details see Odebrecht & Krause 2013 and Odebrecht & Zipser 2013). It should be possible to search for a distinct annotation type or content in several different corpora within the repository. Along with the content requirements the repository needs a structured, machine readable metadata format which can be represented in a graphical interface for the display of information and in the repository system for the different ways to search through the data, e.g. faceted search and free-text search. The meta-model developed from these requirements results in a metadata TEI XML format for the LAUDATIO-Repository but is also designed for and may be applied to other use cases and applications.

The meta-model is designed as an analytic class diagram for which the Unified Modeling Language is used⁴. Such a diagram is useful to document the important issues or concepts in an abstract way. Therefore, the class concepts represent the concepts for the subject-specific application domain 'historical corpora'.⁵

Four main classes are defined: 'corpus', 'document', 'annotationKey' and 'annotationValue' which refer not only to historical corpora but to textual corpora in general:

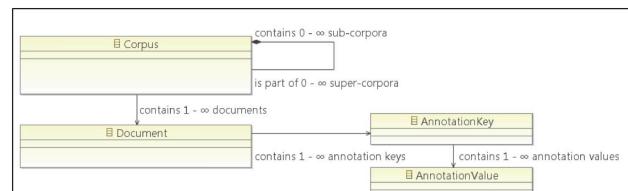


Fig. 1: Meta-model of a corpus. For the sake of concision, the attributes of the classes are left out.

As shown in figure 1, the meta-model⁶ defines a corpus as the sum of all documents regardless of their structure and size. A document is defined as the sum of all annotations regardless of their structure, format and content. 'Annotation' is defined by the sum of all annotation keys and values. For the meta-model, it does not matter whether they have flat, hierarchical

or semantic relations. Every concept carries its own attributes. A 'corpus' is a conceptual collection of digitized and not only linguistically processed (here historical) text. It carries among other things the attributes title, creator, creation date, revision history etc. The class 'document' represents the actual historical text – source text - with its own attributes such as author, date and publication history. The classes 'annotationKey' and 'annotationValue' constitute a document because the sum of transcriptions, normalizations, including segmentations and further annotations build - technically speaking - a 'document'. 'AnnotationKey' in turn carries attributes similar to 'corpus' and 'document' such as date, author and revision history. For example, the attribute 'author' can refer either to the creator of a historical text or to the annotator of a certain annotation layer and may also refer to the same entity or person. This is important for corpus documentation. When re-using corpora, for instance further annotations on an existing corpus are made by third parties, a clear reference can be made to the copyrights. Attributes such as 'date' also refer to every class of the meta-model, meaning that 'corpus' as well as 'annotation' may have a date of creation like 'document' which genuinely has a publication date.

For the technical realization⁷ we used a customization of TEI XML with an ODD specification. The meta-model was mapped to three TEI header structures, one for each concept: a header for 'corpus', 'annotationKey' and 'annotationValue', a header for each 'document' in the corpus and a header for each preparation step of the corpus in general:

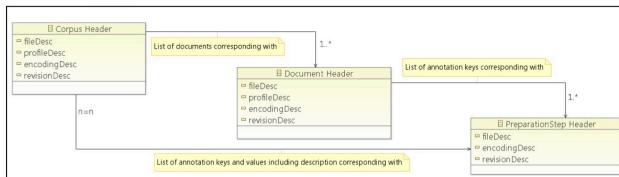


Fig. 2: Technical mapping of the meta-model and the TEI xml header structure

The attributes of each class are mapped into the corresponding TEI element sets. For example, the attributes date and author correspond to the TEI elements <date>, <author> and <editor> with a specifying attribute @role for "annotator" in the <fileDesc> element. The <publicationStmt> element contains the attributes revision and/or publication history for each class. The classes 'annotationKey' and 'annotationValue' are realized with the element set of <elementSpec>. With the help of the attribute @corresp, references between the list of annotation keys and values of the whole corpus to each document and to each preparation step, including information about formats and annotation relations such as segmentations of the annotations, are technically implemented. Each TEI header is customized with the help of ODD⁸. The TEI headers are the technical basis for the uniform display and search of every class and its attributes in the repository. For every corpus, e.g. RIDGES and GerManC, the values for <author> referring to either a distinct annotation layer or a distinct document can be uniformly searched via a faceted search or can be displayed in the corpus view.

The meta-model presented here provides a generic mechanism for the representation of multiply annotated corpora that probably goes beyond the scope of historical corpora alone. Our experience with dealing with a variety of available historical resources has shown how flexible and reliable the model can be in this domain, though work remains to be done in dealing with more relational annotation schemes describing disconnected sources such as annotation between documents in the same or in different corpora.

References

1. GerManC is freely available at hdl.handle.net/11022/0000-0000-1D32-8
2. GerManC is also available in the standoff format GATE which maps all annotations of the TEI XML and the CoNLL formats.
3. The RIDGES Herbology Corpus is freely available at hdl.handle.net/11022/0000-0000-1CDB-B
4. OMG (2009) OmG unied modeling languagtem (omg uml), infrastructure: www.omg.org/spec/
5. For the sake of brevity, the attributes of the classes are left out in figure 1.
6. All ODDs are freely available at www.laudatio-repository.org/repository/documentation.
7. For the technical implementation of the header in the basis systems of LAUDATIO-Repository with the help of elastic search & co see www.laudatio-repository.org/repository/technical-documentation/
8. Each ODD is freely available at korpling.german.hu-berlin.de/schemata/laudatio/doc/S6/
- Burnard, Lou, Rahtz, Sebastian** (2004) *RelaxNG with Son of ODD*. Extreme Markup Languages Proceedings 2004. Montréal, Québec.
- Durrell, Martin, Ensslin, Astrid, Bennett, Paul** (2007) *The GerManC project* In Sprache und Datenverarbeitung 31 (2007), pp. 71-80.
- Foth, Kilian A.** (2006) *Eine umfassende Constraint-Dependenz-Grammatik des Deutschen*. Technischer Report. Universität Hamburg. Hamburg.
- Krause, Thomas, Odebrecht, Carolin, Zielke, Dennis** (2013) *Wie kann der Zugriff, die Wiederverwendung und langfristige Speicherung von linguistischen Korpora realisiert werden?* 35. Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft (DGfS). 12.-15.03.2013. Potsdam Germany.
- Krause, Thomas, Lüdeling, Anke, Odebrecht, Carolin, Zeldes, Amir** (2012) *Multiple Tokenization in a Diachronic Corpus*. Exploring Ancient Languages through Corpora Conference (EALC). 14.-16.06.2012. Oslo Norway.
- Odebrecht, Carolin, Zipser, Florian** (2013) *LAUDATIO - Eine Infrastruktur zur linguistischen Analyse historischer Korpora*. DTA-/CLARIN-D Konferenz und -Workshops: Historische Textkorpora für die Geistes- und Sozialwissenschaften. Fragestellungen und Nutzungserspektiven 18.-19.02.2013. Berlin Germany.
- Odebrecht, Carolin, Krause, Thomas** (2013) *Metadata in an Infrastructure for Historical Corpora. SFB 732 Incremental Specification in Context - Colloquium 20.06.2013*. Stuttgart Germany. http://www.uni-stuttgart.de/linguistik/sfb732/files/abstract_odebrechtkrause.pdf
- Schmidt, Thomas, Wörner, Kai** (2009) EXMARaLDA - Creating, analysing and sharing spoken language corpora for pragmatic research. Pragmatics 19/4. pp.565-582.
- Schiller, Anne, Teufel, Simone, Stöckert, Christine, Thielen, Christine** (1999) *Guidelines für das Tagging deutscher Textkorpora mit STTS*. Technischer Report. Universität Stuttgart, Institut für maschinelle Sprachverarbeitung & Universität Tübingen, Seminar für Sprachwissenschaft.
- Zeldes, Amir, Ritz, Julia, Lüdeling, Anke & Chiarcos, Christian** (2009), "ANNIS: A Search Tool for Multi-Layer Annotated Corpora". In: Proceedings of Corpus Linguistics 2009, July 20-23, Liverpool, UK.

A hipersensibilidade do Território – viver entre terra e nuvens

Oliveira, Lídia

lidia@ua.pt
University of Aveiro, Portugal

Baldi, Vania

vbaldi@ua.pt
University of Aveiro, Portugal

Os lugares aumentados multiplexos

O mapa é ele próprio um dispositivo cultural que foi tendo a sua metamorfose do proto-mapa aos mapas flexíveis em suporte digital. O mapa é um interface cultural, político, geoestratégico, clássico entre o sujeito e o território. O mapa permite relacionar agentes e sobrepor diversas camadas de informação: "O modelo estrutural do mapa é composto por três níveis de base: o mapa sintático, o mapa semântico, e o mapa pragmático. Os diversos níveis permitem ao mapa expressar-se como ferramenta de conhecimento, poder e comunicação." (Neves, 2011, p.1). A passagem do paradigma analógico para o paradigma digital trouxe também ao mapa novas identidades e novas potencialidades. O mapa digital passa a ser dinâmico, permitindo usufruir da gestão de dados geográficos e ambientais complexos que os Sistemas de Informação Geográfica (SIG) passaram a permitir. Desde do surgimento da "*geographic information science*" passaram mais de 20 anos (Goodchild, 2012), neste período deu-se uma mudança radical nos nível de conectividade à internet, na computação ubíqua, na miniaturização dos dispositivos de comunicação que permitiu a sua portabilidade. É neste contexto em que os dispositivos móveis conectados em rede e apetrechados de aplicações que usam informação georeferenciada, em que o mapa surge como interface no qual se partilham conteúdos e se estabelecem redes sociais, que interessa pensar a relação entre mapa e território: onde acaba o mapa e começo o território? A terra como interface (Neves, 2011). E onde começo o território e acaba o mapa? Há uma dupla textualidade.

O mapa como plataforma *geomedia* multifuncional e multidimensional (Neves, 2011) abre, num efeito quase paradoxal, a oportunidade de ter o território a desempenhar a função de interface, de *mapa* que remete para o mapa. O território passa a estar embutido de sensores, de etiquetas RFID, etiquetas QR Code, etc., que o tornam num corpo implantado de dispositivos que o convertem em agente. Simbiose entre biológico, ecológico e a codificação, um meio saturado de capacidade computacional (Kang e Cuff, 2005). A ideia do cyborg, com a incorporação da tecnologia no próprio corpo, salta agora para o território – Cyborgização do território (*Cyborgeo*) – o território como organismo cibernetico, corpo implantado de dispositivos que o tornam agente. Os três atores – natureza (espaço), ambiente construído (território) e subjetividade (usuário/lugar) – interatuam tendo como pele o mapa sensitivo e o território sensitivo que permitem ampliar as relações no cotidiano: "O território humano é o espaço povoado de artefatos tecnológicos. (...) É um mundo de tecnologias infiltradas, das tecnologias que, quanto mais poderosas, mais invisíveis." (Firmo e Duarte, 2012, p.71).

O território écran, o próprio território é o elemento desencadeador da interação devido à sobreposição de territorialidades, o mesmo lugar pode ser desdobrado em camadas virtuais customizáveis graças à computação ubíqua, à realidade aumentada, aos interfaces tangíveis, aos *smart objects* (objetos conectados com a internet), *wearable computers* (dispositivos de computação e telecomunicações embbebidos no vestuário), etc.

Nasce e incrementa-se uma nova dinâmica de relacionamento com o território, de possibilidade de antropológicamente nos situarmos, em que o território ganha novas camadas atuantes e em que o usuário se relaciona consigo, com o território e com os outros numa lógica de convergência e participação.

A análise deve ser realizada em três eixos: a representação, a semântica e a participação (Orellana e Ballari, 2009, p.29). Sendo que os LBSM (Location Based Social Media) – "As LBSM create a new kind of visibility and memory about places, persons and activities I argue that they are significant for the subjective assignment of sense top lace." (Fischer, 2008, p.586) – permitem gerar novas dinâmicas de fruição do território, com uso de dispositivos móveis usando GeoCMS (*Geospatial Content Management System*).

O enraizamento da comunicação na territorialidade abre novas fronteiras na cultura da mobilidade, que deixa de ser apenas mobilidade virtual, navegação no ciberespaço, para ser navegação entre terra e nuvens. Sobre a imagem captada da realidade o utilizador pode obter novas camadas de informação sobre esse lugar, camadas de experiência e relação que

não estariam ao seu alcance, criando ambientes mistos aumentados baseados na colaboração – ACME – *Augmented Collaboration in Mixed Environments* (Lucas, 2012).

Trata-se de uma ecologia midiática híbrida (Santaella, 2008) – novas espacialidade / hipercomplexidade cultural e comunicacional dos lugares – memória e esquecimento, individual e coletivo.

A informação geográfica (GIS), a internet das coisas dada pelas *tags* de radiofrequência (RFID) e a linguagem de marcação geográfica (GML – *Geographic Markup Language*), entre outras tecnologias, criam a oportunidade de desenvolvimento de serviços que permitem gerar uma relação afetiva e sensorial com o lugar (sensorização do espaço → sensibilidade do lugar). Os elementos dos lugares passam a ser atores que dinamizam o usufruto e a fruição estética, emotiva, histórica, política e cultural do lugar. A questão central não é a tecnologia que rapidamente se torna obsoleta, a questão central é o *ontos* do lugar, a sua essência, o que o torna particular, ou seja, é a dimensão geocultural e geoemocional que é central – as pessoas têm no essencial uma relação estética com os lugares, gostam dos sítios. Mais do que um cálculo racional é de um cálculo relacional, afetivo que se trata. Deixam-se afetar pelo lugar. Esta sensorialidade é projetada na própria tactilidade dos dispositivos que interatuam com o território.

Cada lugar esconde um conjunto de informações, de relações potenciais, de desafios, mistérios e oportunidades, com os serviços de informação e comunicação georeferenciados abre-se a oportunidade de tornar visível o invisível dos lugares, de tornar patente o latente. Rede de pontos quentes do lugar (*hotspots places*) – rede de geotags que permite transformar/adicionar ao lugar uma camada que está nas *nuvens*, mas se cola ao lugar ampliando o potencial de relação. O lugar como um corpo com história e com histórias.

A comunicação em rede vê a sua dinâmica enriquecida pela percepção do contexto que se passa a ter. O enraizamento da comunicação na territorialidade abre novas fronteiras na cultura da mobilidade, que deixa de ser apenas mobilidade virtual, navegação no ciberespaço, para ser navegação entre terra e nuvens.

O território interface, desafios futuros

A criação de novos serviços em que a componente cultural e social seja central é o desafio principal. O envolvimento das instituições e organizações culturais a participarem na criação de conteúdos culturais de qualidade, que potenciem a riqueza histórica, cultural, social, política e científica de um lugar.

Este âmbito apresentamos 2 projetos desenvolvidos no Departamento de Comunicação e Arte da Universidade de Aveiro, Portugal. O Projeto Lookin que é um serviço para dispositivos móveis, que permite o reconhecimento de edifícios históricos através de um scan à fachada e partir daí o utilizador pode obter mais informação histórica e cultural sobre o edifício, partilhar comentários e imagens na rede social que esse serviço suporta.

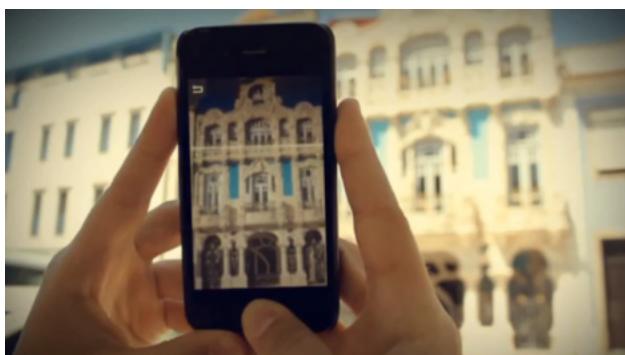


Fig. 1: Projeto Lookin

O outro Projeto, Bioduna , usa informação georeferenciada no contexto de uma reserva natural – Reserva Natural das Dunas de S. Jacinto – para fornecer ao visitante, que faz caminhada nos trilhos da reserva, informação em regime de realidade aumentada. Ou seja, o visitante apontando o dispositivo móvel (telefone, tablete, etc.) à paisagem recebe mais informação sobre a fauna e a flora, nomeadamente, vídeos com os fenómenos que ocorrem noutras épocas do ano. Assim, a sua experiência entre neste território hipersensível, entre terra e nuvens é muito mais rica.

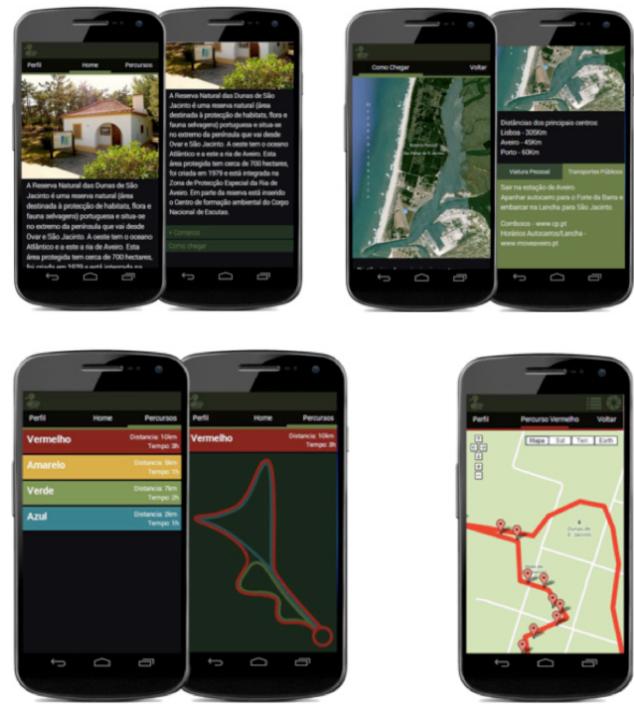


Fig. 2: Projeto Bioduna

Vive-se numa realidade mista – “*Mixed reality defines the sharing of a space-time between the real and the virtual world.*” (Lucasa, Cornishb et al., 2012, p.277) – na confluência da fruição do território com a fruição da territorialidade imaterial da camada cultural e social desse território.

References

- Firmo, R.; Duarte, F. (2012)** *Do Mundo Codificado ao Espaço Ampliado*. In: JANEIRO/FAU/PROARQ, U. F. D. R. D. (Ed.). Qualidade do Lugar e Cultura Contemporânea: controvérsias e ressonâncias em ambientes urbanos. Rio de Janeiro: Afonso Rheingantz e Rosa Pedro. p.69-80.
- Fischer, F. (2008)** *Implications of the usage of mobile collaborative mapping systems for the sense of place*. Real Corp, p. 583-587, 2008.
- Goodchild, M. F. (2012)** *Twenty years of progress: GIScience in 2010*. Journal of Spatial Information Science, n. 1, p. 3-20. ISSN 1948-660X.
- Kang, J.; Cuff, D. (2005)** *Pervasive computing: embedding the public sphere*. Wash. & Lee L. Rev., v. 62, p. 93.
- Lucas, J. F. (2012)** *Interactions et réalité mixte dans la ville hybride*. In: KHALDOUN, Z., HyperUrbain 3 : Villes hybrides et enjeux de l'aménagement des urbanités numériques - Actes de colloque HyperUrbain.3, 2012, Paris. Europa Production.
- Lucasa, J.-F.; Cornishb, T.; Margolisc, T. (2012)** *To a cultural perspective of mixed reality events: a case study of event overflow in operas and concerts in mixed reality*. New Review of Hypermedia and Multimedia, v. 18, n. 4, Special Issue: Cultures in Virtual Worlds, p. 277-293. Disponível em: <<http://www.tandfonline.com/doi/abs/10.1080/13614568.2012.746741> - preview>.
- Neves, P. C. M. A Terra (2011)** como Interface: *O Paradigma do Mapa para o Século XXI*. 113 Dissertação de Mestrado (Mestrado em Ciências da Comunicação). Faculdade de

Ciências Sociais e Humanas, Universidade Nova de Lisboa, Lisboa.

Orellana, D.; Ballari, D. (2009) *La GeoWeb y su evolución: Un marco de análisis en tres dimensiones*. Revista Universidad Verdad, v. 49, n. Agosto, p. 25-52.

MapaHD: Exploring Spanish and Portuguese Speaking DH Communities

Ortega, Élika

eortega@uwo.ca

CulturePlex Lab / University of Western Ontario

Gutiérrez, Silvia

silvia_eunice.gutierrez_de_la_torre@stud-mail.uni-wuerzburg.de
Würzburg Universität

1. Introduction

A community of Digital Humanities in Spanish and Portuguese (HD*) has been consolidating over the past few years. Due to a series of events gathering numerous colleagues and taking place in diverse latitudes, 2013 has been a turning point. A milestone was the first DiaHD, which brought together about a hundred practitioners and showed that HD scholarship is highly active and eager to build a cohesive community. The purpose of the event was "to identify and establish or improve networks and collaborative work among the community of digital humanists in Latin America, the Caribbean, and the Iberian Peninsula, as well as digital humanists in other regions of the world whose work is done in Spanish or Portuguese" (translation ours)¹. Our project, MapaHD originated that day and has embraced the bilingual profile established by that event. MapaHD is an exploration of the features and intersections among those who self-identify as HD practitioners and their characteristics beyond language affiliation. In our paper, we provide insights into issues of temporal development of HD, geographic location, interdisciplinary practices and approaches, and how progressively a community of digital humanists has been taking shape. The development of MapaHD has been made public from its beginnings at mapahd.org where we have gathered visualizations and preliminary results. Simultaneously, the data collected has been used to build an interactive and exploratory map using DARIAH-DE Geo-Browser that is also available through our website.

1.1. Overview

MapaHD is a direct address to the question launched by Domenico Fornante, "Is there a non Anglo-American Digital Humanities, and if so, what are its characteristics?"². In this project we have gathered and analyzed practitioners' data that evidences not only the existence of a thriving DH community in Spanish and Portuguese languages, but more importantly what its features are. The diversity of the characteristics we have observed sheds light on the HD community's institutional and project affiliations, area of research, geographic location, research approaches, and temporal data.

1.2. Methodology

In order to tackle these issues, we have carried out three research phases:

1) Data collection gathered entirely online through a survey, answered voluntarily by 85 participants. The survey was available during a four-month period from June 10th to October 10th, 2013. Questions included gathered data on

participants' institutional, project, and disciplinary affiliations, research approaches, location, among others. The survey was distributed through mailing lists and Twitter. Links to the projects' survey were tweeted using hashtags used by similar events and communities such as #DíaHD, #HDH2013, #DH2013, #ThatCampBaires, #HDBr, #RedHD, #aroundddh, #HumanidadesDigitales, and #dhpoco. The aim of this distribution model was to catch the attention of as many participants in as many locations as possible. This approach to data collection sought to allow anybody who identified himself/herself as a digital humanist in the two languages to self-report their characteristics, rather than send out invitations to those we might consider to fall even under a "big tent" definition of digital humanities.³

2) Using the available data, we built a graph database organized according to the semantic network schema in Fig 1. Rather than looking for person to person connections, this analysis sought to shed light on non-obvious and non-personal connections among participants. The links joining researchers and students among them are disciplines, approaches, work spaces, and geographic proximity. The data was subjected to frequency, central tendency, and network analysis. Several results emerged from the data contained in the database and are presented in the next section.

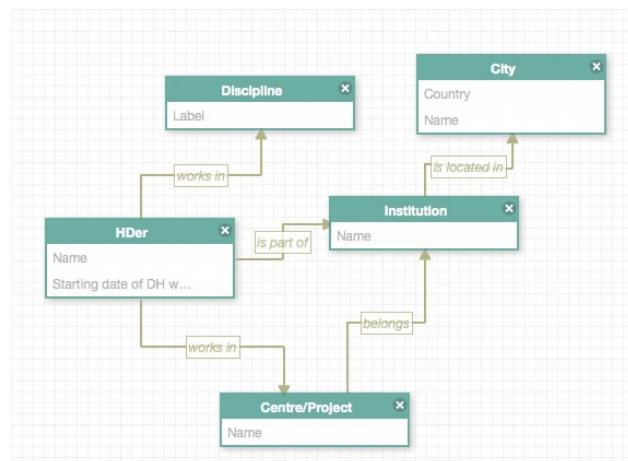


Fig. 1: Database Schema.

3) An interactive map visualization built on DARIAH-DE Geobrowser. This resource is a second contribution of this project to the field as it seeks to serve not only as a visualization of the data collected, but also as a reference tool. Finally, aside from providing a glimpse of the state of the field, MapaHD hopes to strengthen and expand the sense of community and connection among Spanish and Portuguese speaking digital humanists initiated by DiaHD.

2. Results

2.1. Discipline Outlook

More than half of the 85 participants reported working in at least two disciplines distributed as shown in Fig 2. Literary studies was the discipline most participants reported. However, out of the 56 participants who reported working on the literary fields, 32 also reported working in other disciplines.

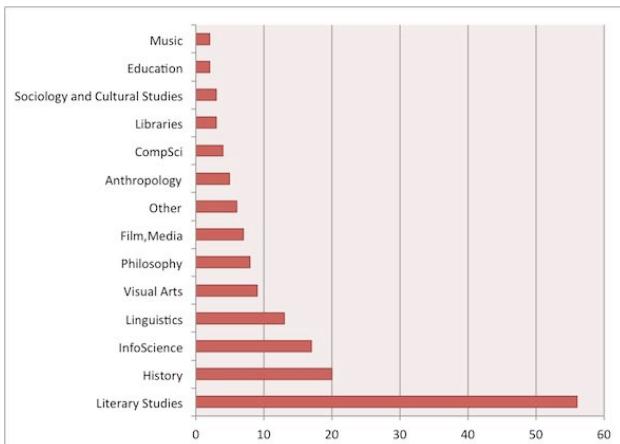


Fig. 2: Discipline distribution of participants.

In Fig 3 we show the most common combinations of literary studies and other disciplines. Aside from the recurrence of literary studies together with other disciplines, the only other recurrent disciplinary combination was History and Visual Art. The rest were mostly unique combinations. This information confirms the fact that, not unlike DH, HD is also “a hybrid domain, crossing disciplinary boundaries and also traditional barriers between theory and practice, technological implementation and scholarly reflection”⁴.

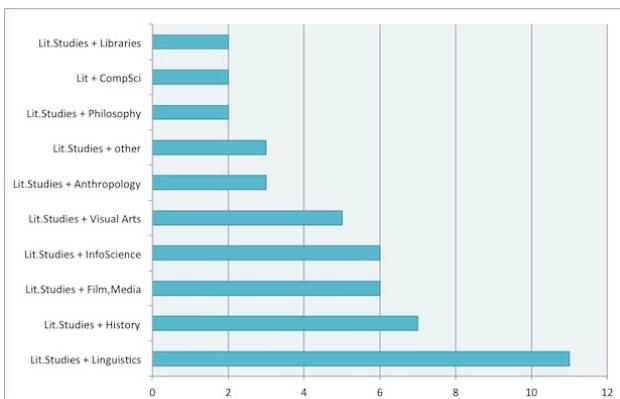


Fig. 3: Combinations between Literary Studies, the most common discipline in the database, and other fields.

Our analyses offer qualitative insights as to what different disciplines might be bringing in into the mix. For example, the recurrence of interdisciplinary work attached to a “root” field of Literary studies suggests that the field is not necessarily the gravity centre of HD as has been suggested by Azofra⁵ but, as observed in Fig 4, a hub where other expertises converge, shedding light upon each other. In contrast, network analysis has shown that the second and third best connected disciplines are Information Sciences and History.

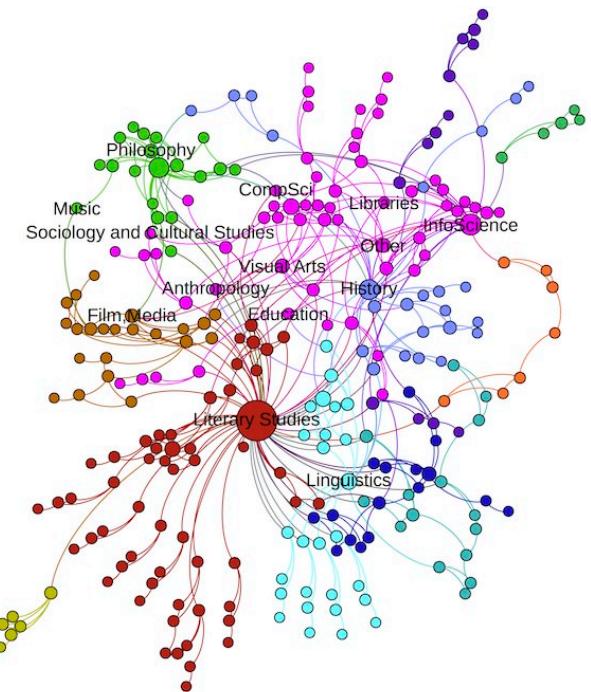


Fig. 4: Network visualization showing the prominence of Literary Studies by measuring degree, and History and Information Science as the next two best connected disciplines.

The relevance of these two fields in the network resides not so much in how many participants reported them, but how variably combined they are. As a matter of fact, the cluster formed around Information Sciences brings together a few other disciplines such as Education and Computer Science. Even though they have key links joining them to the rest of the network, disciplines such as Philosophy, Film and Media Studies, and Linguistics retain a certain level of isolation. From our data it is possible to see both the exposing of disciplines to other fields of knowledge, on the one hand; and on the other, how in opening up, disciplines flood other fields too. While some disciplines, by sheer numbers seem to be exercising a larger influence, connecting fields like History and Information Sciences might be providing a common foundation of porous perspectives through which these dynamics take place.

2.2. Geographic Outlook

The geographic location of MapaHD participants (Fig 5) was a foundation for the initial concept of the project. Through the interactive map visualization, we also explored the participants distribution. In total 41 cities located in 11 countries were identified (Fig 6). As a reference, this is close to 50% of the total number of countries represented in ACH membership set at 23, as Bethany Nowviskie reported via Twitter in October 2013⁶.



Fig. 5: Screenshot of MapaHD, built on DARIA-DE Geobrowser, showing the spread of the HD community around the world.

Interestingly, close to half of the locations are found in the UK, USA, Canada, Germany, and Italy where, though common, neither Spanish nor Portuguese are official languages. Although some of the participants’ location can be seen as a ‘diaspora’,

our results have showed that this is only a small portion of them (22%) and not the sole distinctive of the analyzed group.

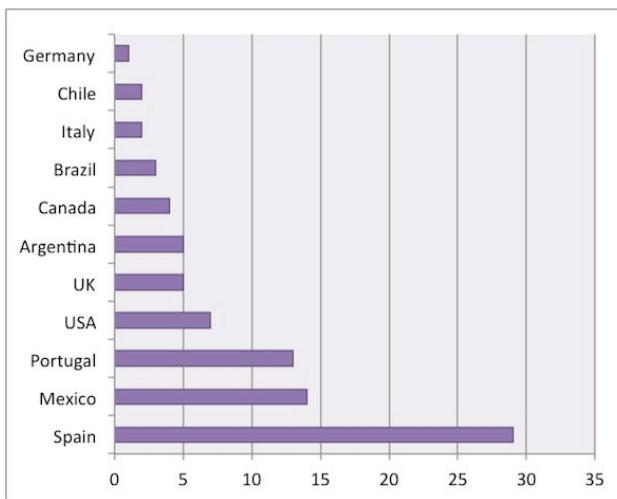


Fig. 6: Geographic distribution of participants.

The issue of location and problems of centrality and periphery in the context of Digital Humanities have been better expressed by Domenico Fiornonte, who stresses the fact that DH have not "succeeded in either strengthening the field of humanities or putting some balance into the power relationships between humanities and computer science" ⁷. We believe that, as Isabel Galina proposes, this lack of balance "can also help us think about DH from a different perspective... [and] pushes the limits of our creativity and our capacity to solve problems" ⁸.

The clear ties with the Anglo-American branch of DH, and to a lesser extent with the continental European one, seem to imply that as an identifiable community HD is porous and prone to cross-pollination in terms of approaches, academic practices, and language. The international spread of HD practitioners might be the cause behind the particular diversity in this community. Nevertheless, the HD community maintains a sense of cohesion that can be traced perhaps to the shared lack of visibility, institutional similarities, and linguistic coincidences – Rafael Alvarado's "network of family resemblances" ⁹. Furthermore, HD ties with other DH communities, both geographic and linguistic, have set up a communication channel through which approaches and projects may travel back and forth.

3. Conclusions

Much has been said about the characteristics of DH on a global scale and, especially, of the Anglo-American branch. MapaHD constitutes the first data-driven approach to the HD community and we provide insights into its disciplinary and geographical particularities. Not discussed in this abstract but included in the paper are the issues of collaboration and where it takes place, as well as an outlook of the connections among the research approaches undertaken, and the historical development of our participants' trajectories.

Endnotes

*We use HD to distinguish the Spanish and Portuguese speaking branch of digital humanities from the Anglo-American one commonly referred to as DH.

References

1. "Día de las Humanidades Digitales 2013. Día HD 2013. Día de las Humanidades Digitales. 10 May 2013. Web. 25 August 2013. Par. 1 dhd2013.filos.unam.mx/acerca/
2. Fiornonte, Domenico (2012). *Towards a Cultural Critique of the Digital Humanities*. Historical Social Research. Historische Sozialforschung. Vol. 37. Print. p. 59.
3. Azofra, Elena. *Humanidades digitales cerca del 'Finis terrae'*. MorFlog. 9 July 2013. Web. 13 Sept. 2013. <http://morflog.hypotheses.org>
4. Flanders, Julia, Wendell Piez, and Melissa Terras (2013). *Welcome to Digital Humanities Quarterly*. DHQ: Digital Humanities Quarterly: The Alliance of Digital Humanities Organizations, 2007. Web. 24 Sept. Par. 3 www.digitalhumanities.org/dhq/vol/001/1/000007/000007.html
5. Nowviskie, Bethany (nowviskie). *Our numbers have grown & diversified since this page was last updated: ach.org/membership/ -- 480 ACHers now hail from 23 countries*. October 14th, 2013, 8:50 a.m. Tweet.
6. Galina, Isabel. *Is There Anybody Out There? Building a Global Digital Humanities Community*. Red HD. Red de Humanistas Digitales. 19 July 2013. Web. 4 Sept. 2013. Par. 16 humanidadesdigitales.net/blog/2013/07/19/is-there-anybody-out-there-building-a-global-digital-humanities-community
7. Fiornonte, Domenico (2012). *Towards a Cultural Critique of the Digital Humanities*. Historical Social Research. Historische Sozialforschung. Vol. 37. Print. p. 72.
8. Galina, Isabel. *Is There Anybody Out There? Building a Global Digital Humanities Community*. Red HD. Red de Humanistas Digitales. 19 July 2013. Web. 4 Sept. 2013. Par. 16 humanidadesdigitales.net/blog/2013/07/19/is-there-anybody-out-there-building-a-global-digital-humanities-community
9. Flanders, Julia, Wendell Piez, and Melissa Terras (2007). *Welcome to Digital Humanities Quarterly*. DHQ: Digital Humanities Quarterly: The Alliance of Digital Humanities Organizations. Web. 24 Sept. 2013. Par. 3 www.digitalhumanities.org/dhq/vol/001/1/000007/000007.html

Geoweb 2.0 and Design Empowerment: A Critical Evaluation of Eleven Cases

Pak, Burak

burak.pak@kuleuven.be
KU Leuven Faculty of Architecture

Verbeke, Johan

johan.verbeke@kuleuven.be
KU Leuven Faculty of Architecture

1. Introduction

The concept of public participation was brought onto the agenda of urban design and planning prominently after the May events of 1968 (Jencks, 2011). Arnstein (1969)¹ was the first to identify various ways of participation: *manipulation, therapy, informing, consultation, placation, partnership, delegated power and citizen control*. After this study, it became more evident that facilitating participation practices do not necessarily grant planning power to the citizens; they may manipulate them as well.

Following the Arnstein's ladder, the understanding of participation shifted towards the greater democratization of the processes and deeper involvement of citizens. Connor (1988)², Dorcey et al. (1994)³ and Rocha (1997)⁴ have proposed their updated versions of the participation ladder, each focusing on slightly different aspects. Connor (1988)'s point of view was oriented more towards conflict resolution whereas Dorcey et al. (1994) suggested ongoing involvement and consensus building as the highest level of participation. Rocha (1997) placed political empowerment at the top and atomic empowerment at the bottom of her version of the participation ladder.

Senbel and Church (2011)⁵ linked various forms of empowerment and visualization media while proposing a more "enabling" version of Arnstein's ladder. Their ladder involved

six "instances" of **design empowerment**. The highest level on this ladder is **independent design, when ordinary citizens gain the capacity to create their own plans and visions; reaching autonomy.**

Overall, the brief review above illustrates the theoretical shift or the "communicative turn" from rational planning to deliberative planning.

From the perspective of **geospatial participatory technologies**, it is possible to track similar layers of transformation regarding the production and dissemination of geographic information. From top-down to bottom-up, referring to the public participation GIS (PPGIS), from "requested production" to "voluntary production", and finally, towards the wikification of GIS and **Web 2.0-based social-geographic applications (Geoweb 2.0)**(Roche et. al., 2012)⁶.

Relying on a combination of social software and information aggregation services, Geoweb 2.0-based participatory planning practices stand as a strong alternative to the traditional linear and hierarchical knowledge production methods. These are loaded with constructivist learning and production principles embedded in the ways they enable social knowledge construction (Pak and Verbeke, 2012)⁷.

In this context, we would like to critically address the following questions in our study:

- Which inclusion strategies and tools are used for design empowerment in popular Geoweb 2.0 supported participatory planning practices?
- To what extent do these practices facilitate participation in urban planning?

Motivated with the questions above, we made an evaluation of relevant practices through an online survey. We will share the method and results of this survey and discuss our findings in Section 2. This discussion will be followed by the conclusion (Section 3), in which we summarize the findings and discuss their possible implications for future developments.

2. Evaluation of Design Empowerment Strategies Employed in Practice

We grounded our survey on Senbel and Church's (2011) theoretical framework for "**design empowerment**". As briefly described in the introduction, the authors proposed six instances of citizen involvement in design (Figure 1). In this framework, *independent design* is depicted as the highest level of empowerment, followed by *integration* which involves the coproduction of plans. *Inclusion* of the thoughts of the participants among other priorities, *ideation* about the plans and *inspiration* triggering response to an alternative and *information* are the relatively lower instances of design empowerment.

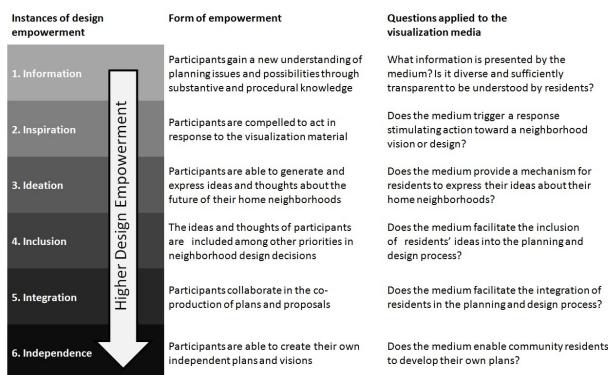


Fig. 1: Senbel and Church's (2011) Instances of Design Empowerment

Based on the instances and forms of design empowerment above, we prepared an online survey to analyze the inclusion strategies and tools used for design empowerment in the existing Geoweb 2.0 supported practices. In January 2013, we distributed the survey to thirty organizations listed by the crowdsourcing.org⁸ directory as related to urban design and planning. These organizations were contacted via three

different communication channels: email, phone and their facebook pages.

Eleven organizations have accepted to attend our survey (*OpenPlans*, *Nextdoor*, *CitySourced*, *Neighborhood*, *LocalWiki*, *Spacehive*, *MindMixer*, *LocalData*, *mySociety*, *Ideavibes*, *CommunityPlanIt*). At the time of the survey, these organizations represented dominant North America and UK-based practices which operate globally, including the Continental Europe. 64 percent of the participants were private organizations. The remaining 36 percent were NGOs, Social enterprises and University laboratories.

In relation to Senbel and Church (2011)'s instances of design empowerment, we asked the participants to rank the priorities (Table 1) of their practices. A legal representative of each organization answered our survey.

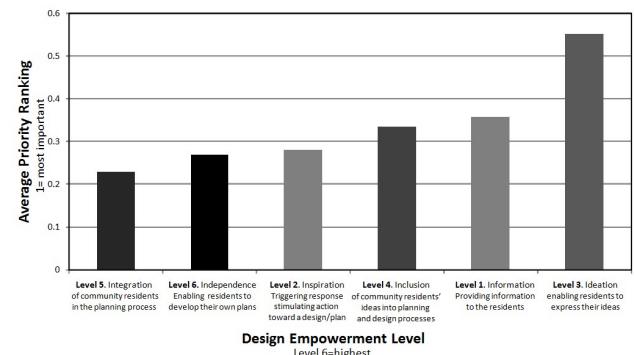


Fig. 2: Average rankings of the priorities of the Geoweb 2.0 applications (higher levels of design empowerment are indicated as darker colors)

According to the participants, *Ideation* (empowerment level 3) was the most important priority, followed by *Information* (level 1) and *Inclusion* (level 4). The two highest levels of empowerment -*Independence* and *Integration*- were ranked as the two least important priorities (Figure 2).

Following the ranking, six of the participants chose to answer an open question on the design empowerment potentials of their Geoweb 2.0 applications. One participant expressed that their application "*can be leveraged by people trying citizen design*". According to another participant, the organization "*had lots of broad efforts around planning, driven entirely by citizens. But it had little official use by city planners or professional planners*".

One of the other participants indicated that their "*toolkit is less about formulating citizen-designed plans, but it rather provides a more efficient method for data collection already taking place*". Similar to this comment, another wrote that their application was "*designed more for reporting problems with the local area (e.g. potholes, broken street lights) than for any integration with urban planning*".

In addition to the observations above, we made a brief analysis of the provided functions (Table 1). The most common ones were: commenting on other users' contents (91 percent), followed by adding a placemark and descriptive text (82), tagging content based on predefined categories (64) and uploading a document (55 percent).

Only one of the Geoweb 2.0 applications supported annotated drafting and drawing tools, which are necessary for the empowerment of citizens in the independent and collaborative design of plans and projects.

Provided Functions	Percentage
Commenting on other users' contents	91%
Adding a placemark and descriptive text	82%
Tagging content based on predefined categories	64%
Uploading a document	55%
Adding a geolocated photo	45%
Editing other users' contents	45%
Tagging content based on user-defined categories	55%
Forum	36%
Internal Messaging	27%
User controlled thematic layers	18%
Timeline	18%
Other: Search, Video, Organizer moderation, Civic profile, Email notifications, shapefile/kml; data management; survey creation	18%
Drawing polygons on the map and adding a description	9%

The last finding was on the intended target audience of the Geoweb applications. According to the participants these were Neighborhood Organizations (100 percent), Community Residents/inhabitants (100 percent), Governmental Administrations (91 percent), Governmental Planning Organizations, NGOs and Umbrella Organizations, Urban Designers, Research Organizations (73 percent), Property and Land Owners and Project, Real Estate Developers (55 percent), Architects (36 percent), Others (18 percent) and Financers (9 percent).

3. Conclusion and Discussion

Our analysis results suggest that the strategic positioning of the sampled set of Geoweb 2.0 applications was less towards higher levels of design empowerment and more towards **data collection, information and ideation**. This finding was evident in the individual rankings of empowerment intentions as well as the provided tools and functions.

As a response to the open question, participants reported a specific scenario in which the authorities and experts were empowered through the collection of information from the citizens. The intended levels of design empowerment of the citizens were indirect and limited.

Only 9 percent of the practices provided drafting-drawing tools which are evidently essential for the citizens to create their own plans/visions and reach autonomy.

When combined with the self-reported target audiences, our findings suggest that the sampled Geoweb 2.0 applications were primarily intended to be used as **a single-sided communication channel between the citizens and the planning organizations**. None of them included convincing mechanisms to guarantee the consideration of the data collected from the citizens and the inclusion of these into the design and planning processes.

Furthermore, according to the survey results, the majority of the practices (64 percent) were controlled by private organizations. Reflecting on the negative experiences of Facebook and Google (Bucher, 2012)⁹ and (Habermas, 2006)¹⁰ we can claim that public opinion on urban planning cannot be formed in a truly democratic manner without separation of tax-based state from market-based society. Unregulated private social networks may encourage disempowerment due to the commodification of personal and sensitive information on citizens, triggering counter results.

Therefore, for better practices in the future, it is of utmost importance to construct self-regulating and independent systems which can:

- Operate as a mediating interface between the planning authorities and the society,
- Enable inclusion and equal opportunity for participation in innovative ways,
- Ensure the privacy and security of the participants,
- Mobilize discussion on relevant topics and claims and planning actions,
- Promote critical evaluation from different perspectives.

In this context, the potentials of Geoweb 2.0 to empower ordinary citizens to develop their own plans are yet to be harnessed. Reporting potholes can raise awareness but is only a small step in the empowerment ladder.

Acknowledgements

This paper is partially based on post-doctoral research project supported by INNOVIRIS, The Brussels Institute for the encouragement of Scientific Research and Innovation.

References

1. Arnstein, S. R. (1969), *A Ladder of Citizen Participation*. In Journal of the American Planning Association 35 (4). 216–224. Routledge: New York.
2. Connor, D. M. (1988), *A New Ladder of Citizen Participation*. In National Civic Review 77(3), 249-57. National Civic League: Jossey-Bass.
3. Dorcey, A. H. J. (1994), *British Columbia Round Table on the Environment and the Economy. Public involvement in government decision making: choosing the right model*. The Round Table: Victoria, British Columbia.
4. Rocha, E. (1997), *A ladder of empowerment*. Journal of Planning Education and Research, 17(1), pp.31–44. DOI:10.1177/0739456X9701700104
5. Senbel, M., Church, P. S. (2011), *Design Empowerment: The Limits of Accessible Visualization Media in Neighborhood Densification*, in Journal of Planning Education and Research 31(4), 423–437.
6. Roche, S., Mericskay, B., Batita, W., Bach, M., Rondeau, M. (2012), *WikiGIS Basic Concepts: Web 2.0 for Geospatial Collaboration*, in Future Internet 2012, 4, 265-284. Basel: MDPI Publishing.
7. Pak, B., Verbeke, J. (2012), *A Web-based Geographic Virtual Environment for the Analysis and Evaluation of Alternative Urban Development Projects Prepared for Brussels*, ACADIA 2012 Annual International Conference Synthetic Digital Ecologies October 18-21,
8. Crowdsourcing.org (2013), Available from: crowdsourcing.org/ Open Source Repository (accessed 15 October 2013).
9. Bucher, T. (2012), *Want to be on the top? Algorithmic power and the threat of invisibility on Facebook*, New Media & Society November 2012 vol. 14 no. 7 pp. 1164-1180.
10. Habermas, J. (2006), *Political Communication in Media Society: Does Democracy Still Enjoy an Epistemic Dimension? The Impact of Normative Theory on Empirical Research*, Communication Theory 16 (2006) pp. 411–426

A vocabulary of the aesthetic experience for modern dance archives

Paquette-Bigras, Ève

Université de Montréal, Canada

Forest, Dominic

Université de Montréal, Canada

Introduction

This research falls within the field of digital humanities; the arts and information science engage in dialogue. In the last few decades, dance has become a distinct research subject. Dance research needs data. Dance performances remain elusive, and the traces they leave in their wake need to be documented. However, the documentation practices of performance remain unsatisfactory (Couch 1994: 42; Rowat 2005; Desalme 2007: 13; Chaffee 2011: 125), and the specificity of performing arts such as dance is rarely and quite imperfectly taken into account (Le Boeuf 2002; Miller and Le Boeuf 2005).

Dance description in archives needs to be improved because, in this era of massive digital information, the quality of the description impinges on access to the documentation. Description is entangled with access. The better the description, the more efficient the access will be for information seekers. As Nena Couch (2004: 53) once said about dance collections, "If there is no standard language through which a patron may communicate his or her search, the material may be as lost as if the library had never acquired it." Knowledge extraction seems to offer new opportunities in this regard.

Objectives

The goal of this research is to contribute to the development of information management tools by evaluating the relevance of knowledge extraction in information resources maintenance and development for performing arts such as dance. Aesthetic experience is an essential part of art; a part, however, that is hard to define, let alone describe, in an archives context. Through knowledge extraction, we obtain a vocabulary for describing the aesthetic experience of modern dance in archives. Choreographic works were described using this vocabulary.

Methodology

Many contemporary artists include archival material in their artistic practices (Poinsot 2004; Lemay 2009). Performing arts archives and archivists must be as creative as the art they keep records of (Johnson and Fuller Snyder 1999; Jones, Abbott and Ross 2009: 166; Chaffee 2011: 125). Artists are inspired by archives, and archivists should be inspired by artists in return.

To obtain a vocabulary for representing the aesthetic experience of modern dance, we drew on modern literature and modern art. At the end of the 19th century, French writer Stéphane Mallarmé praised in *Autre étude de danse* the dazzling performances of Loïe Fuller, a pioneer of modern dance. Mallarmé was an artist and an aesthete (Delfel 1951), a "métaphysicien du ballet" (Levinson 1983). His work is an aesthetic experience in itself, and is related to dance in many ways (Richard 1961; Kristeva 1974: 537; Block 1977: 96; Levinson 1983; Zachmann 2001). Writing about Mallarmé, Mary Ann Caws (1998: 86) says, "[w]hat he gives us is everything that comes after him." Gayle Zachmann (2001: 188) mentions that "writers and critics [...] have highlighted this poet's contributions to the theoretical underpinnings and reading of modern dance and/or the significance of his writings on dance for his own aesthetic." His work foreshadowed the spring of modern dance.

We worked from a corpus of texts that includes Mallarmé's collections *Divagations* (1897 edition published by Eugène Fasquelle) and *Poésies* (1899 edition published by Émile Deman) for a total of 119 documents (comprising poetic prose, poem, dialogue), 11,850 types on 70,507 tokens (before lexical filtering) and 7,238 types (after lexical filtering). This corpus has been linked to modern dance for decades, and artists as well as experts in the field of dance studies have drawn on it for inspiration.

The vocabulary was obtained through knowledge extraction methods; to be specific, text mining algorithms combining term extraction and clustering. Documents were grouped into clusters using discriminant features (terms). Clustering is the method of choice for thematic discovery and terminology

building (Ibekwe-SanJuan 2007; Forest 2012) and, as such, a bottom-up hierarchical clustering method was used. Once the cluster structure was created, characteristic terms were extracted from each cluster to use as the basis for building a basic structured vocabulary of the aesthetic experience of modern dance.

Results

Two main clusters emerged from the corpus, one of 75 documents, the other of 34 documents. From the 75-document class were drawn three qualifiers, one for each subcluster, of the aesthetic experience of modern dance: *petit* (small), *seul* (in solo), *beau* (beautiful). The antonyms were then drawn directly from the corpus: *grandiose* (grandiose, great), *en couple* (in duo) or *en troupe* (in a body), *laid* (ugly). From the 34-document class were drawn nine topics related to the aesthetic experience: *corps* (body), *idéal* (ideal), *nature* (nature), *nudité* (nudity), *pureté* (purity), *rire* (laughter), *solitude* (solitude, loneliness), *temps* (time), *voix* (voice). The terms allow the description of what is happening visually onstage beyond the storyline, that is, the visual experience.

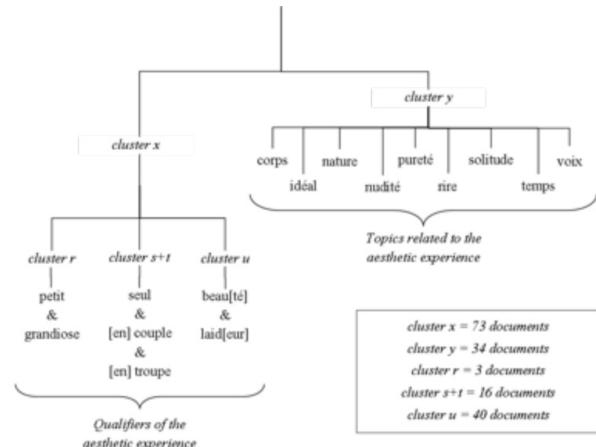


Fig. 1: Vocabulary of the aesthetic experience extracted from Mallarmé's work

We were able to describe choreographic works with the vocabulary. For example, we described Dave St-Pierre's *Un peu de tendresse, bordel de merde!* using the following terms: *troupe*, *voix*, *nudité*, *rire*. St-Pierre's work is known for theatrical staging, shocking nudity and dark humour.

Exploring Mallarmé's vocabulary has allowed us to better describe dance performance and to develop minimal yet innovative access points to traditional archives. Thus this initial experiment supports the relevance of knowledge extraction in information resources maintenance and development for performing arts such as dance. Knowledge extraction is one of many solutions for creating a vocabulary for dance archives.

Conclusion

In this research, we delved into the core of an art, literature, to find a vocabulary for describing art. Humanities computing and non-computing approaches are complementary here. The field of digital humanities, recent advances in information technology and opportunities offered by knowledge extraction all contribute to the possibility of exploring innovative solutions to improve the description of dance performance in archives, as well as fostering a better understanding of the art of dance. In this way, information science empowers the arts while the arts empower information science.

References

- Block, H.** (1977). *Mallarmé and the Symbolist Drama*. Westport: Greenwood Press.
- Caws, M.** (1998). *Mallarmé's Progeny*. In Cohn, R. and Gillespie, G. (eds), *Mallarmé in the twentieth century*. Cranbury, London, Mississauga: Associated University Press, pp. 86-91.
- Chaffee, G.** (2011). *Preserving Transience: Ballet and Modern Dance Archives*, Libri: International Journal of Libraries and Information Services, 61(2): 125-130.
- Couch, N.** (1994). *Dance Collections*. In Sheehy, C. (ed), *Managing Performing Arts Collections in Academic and Public Libraries*. Westport: Greenwood Press, pp. 41-72.
- Delfel, G.** (1951). *L'esthétique de Stéphane Mallarmé*. Paris: Flammarion.
- Desalme, A.** (2007). *Sur les pas de la danse: l'exemple des bibliothèques américaines*, Bulletin des bibliothèques de France, 52(4): 13-22.
- Forest, D.** (2012). *Fouille de textes et analyse thématique: techniques prédictive et descriptive pour la gestion et l'analyse de documents textuels non structurés*. Sarrebruck: Éditions universitaires européennes.
- Ibekwe-SanJuan, F.** (2007). *Fouille de textes: méthodes, outils et applications*. Paris: Hermès, Lavoisier.
- Johnson, C. and Fuller Snyder, A.** (1999). *Securing our Dance Heritage: Issues in the Documentation and Preservation of Dance*. Washington: Council on Library and Information Resources.
- Jones, S., Abbott, D. and Ross, S.** (2009). *Redefining the Performing Arts Archive*, Archival Science, 9(3-4): 165-171.
- Kristeva, J.** (1974). *La révolution du langage poétique: l'avant-garde à la fin du XIX^e siècle, Lautréamont et Mallarmé*. Paris: Éditions du Seuil.
- Le Boeuf, P.** (2002). *Le spectacle vivant en tant qu'objet documentaire et le modèle conceptuel de données des FRBR*. In Arts du spectacle: patrimoine et documentation, XXIII^e congrès international, Paris 25-30 septembre 2000. Paris: Société internationale des bibliothèques et musées des arts du spectacle, Bibliothèque nationale de France.
- LeMay, Y.** (2009). *Art et archives: une perspective archivistique*, Encontros Bibli, édition spéciale: 64-86.
- Levinson, A.** (1983). *The Idea of Dance: From Aristotle to Mallarmé*. In Copeland, R. and Cohen, M. (eds), *What Is Dance?* New York: Oxford University Press, pp. 47-54.
- Mallarmé, S.** (1897). *Divagations*. Paris: Bibliothèque Charpentier, Eugène Fasquelle. <http://fr.wikisource.org/wiki/Divagations> (accessed May 2013).
- Mallarmé, S.** (1899). *Poésies*. Bruxelles: Edmond Deman. http://fr.wikisource.org/wiki/Po%C3%A9sies_%28Mallarm%C3%A9%29 (accessed May 2013).
- Miller, D. and Le Boeuf, P.** (2005). "Such stuff as dreams are made on". How Does FRBR Fit Performing Arts, Cataloging & Classification Quarterly, 39 (3/4): 151-178.
- Poinset, J.-M.** (2004). *Avant-propos*. In Mokhtari, S. (ed), *Les artistes contemporains et l'archive*. Rennes : Presses universitaires de Rennes, pp. 5-8.
- Richard, J.-P.** (1961). *L'univers imaginaire de Mallarmé*. Paris: Editions du Seuil.
- Rowat, T.** (2005). *Review of Recent Approaches to Growing Dance Legacy*. Ottawa: Canada Council for the Arts.
- Valaskakis-Tembeck, I., Odom, S. and Fisher-Stitt, N.** (1997). *Dance Research in Canada*, Dance Research Journal, 29(1): 107-110.
- Zachmann, G.** (2001). *Offensive Moves in Mallarmé: Dancing with des astres*. In Grossman, K., Lane, M., Monicat, B. and Silverman, W. (eds), *Confrontations: Politics and Aesthetics in Nineteenth-Century France*. Amsterdam, Atlanta: Rodopi, pp. 187-200.

Mixing contributions, collaborations and co-creation: participatory archaeology through crowd-sourcing

Pett, Daniel Edward John

The British Museum, United Kingdom

Bonacchi, Chiara

The Institute of Archaeology, University College London, United Kingdom

Bevan, Andy

The Institute of Archaeology, University College London, United Kingdom



Micro Pasts

This paper reflects on the pros and cons of a mixed contributory, collaborative and co-creative model of community engagement (Bonney et al. 2009) with archaeology through crowdsourcing and crowd-funding. It arises from a recently initiated project entitled Crowd- and Community-fuelled Archaeological Research. This project is collaboration between University College London's Institute of Archaeology and The British Museum and is codenamed MicroPasts. This project is developing a web platform to promote new online communities that span hitherto different kinds of archaeological enthusiasts, from 'traditional academics', to already established groups of volunteer interest such as archaeological societies, and a wider crowd of potential contributors.

These potentially diverse and international communities will have the opportunity of collaborating on one or more of three things:

- 1. Co-production of open licensed research data, such as 3D models of Bronze Age metal objects from England, or the tagging of historical photographs of early 20th century archaeological excavations in the Levant;
- 2. Collaborative development of completely new research projects, where several different kinds of contributor will be involved. The intellectual lead need not be an institutionally affiliated academic;
- 3. Crowd-funding of some of the latter projects alongside a number of other community archaeology initiatives. We hope to thereby formalise and build upon some pioneering but rather spontaneous experiments using crowdsourcing to co-design new research (e.g. Old Weather, or Herbaria@home; Ridge 2013). For example, we will begin with 'scaffolded' contributory activities where members of the public are invited to participate in research agendas already outlined by academics. It is then hoped that the contributors can develop these research agendas into a higher or even tangential direction.

On the MicroPasts platform, we hope to foster an ever increasing sequence of participation, by encouraging those involved in data co-production via standard crowd-sourcing to participate in follow-up co-design and also to perhaps fund their projects through crowd-funding initiatives. MicroPasts contributors are entirely free to focus their efforts on just one, or if they prefer, several of these areas. To our knowledge, this is the first model of participatory community engagement of this

kind in the archaeology and cultural heritage sector. This paper will present both this overall project rationale and the technical choices we have made to turn it into reality.

All the software employed in this project will be open source and any modifications we make will be released to the open source community via an institutional GitHub account. For example, we will be using a British Museum hosted multi-site Wordpress installation for our main portal and for blogging about our research. We have already customised an instance of the CrowdCrafting platform (see Mansell 2012:8) and the first application to go live on this will allow transcription of the British Museum's New Bronze Age II card index. For community interaction, we are using the Discourse platform for discussion and help to be dispensed by the contributors and facilitators. For the crowd-funding element of the project, we will be implementing an instance of the Neighborly community fund raising software (based on the successful Brazilian Catarse software.)

We will demonstrate how these software packages contribute to the holistic model of participatory archaeological research that we have in mind especially with regard to how the challenging task of effective co-design might be achieved. We will also discuss how different crowd-sourcing tasks have been incorporated into the CrowdCrafting platform, the challenges and obstacles we faced during this process (for example institutional opposition to micropayments), what influenced the choices that we made and how different tasks succeeded or failed. In so doing we hope to provide wider inspiration for others in the fields associated with Digital Humanities and a replicable model for anyone to copy. For instance, all the software created for this project could conceivably be reused by anyone with an interest in crowdsourcing or crowdfunding.

Finally, we will consider what kind of evaluative framework is suitable for understanding community engagement in such activities. We are interested in understanding the experiential, cultural and economic values behind contributors' involvement in crowdsourcing and crowd-funding research into the human past. To achieve this aim, the following areas will be investigated individually and their inter-relational synthesis:

- (a) Motivations for contributing;
- (b) Dynamics of community building and the kinds of relationships that are built and sustained;
- (c) Cultural and economic resources mobilised via community members.

Our analysis will address existing works on these topics within the science and cultural heritage domains, but will also look to deepen their insights with respect to archaeology and history in particular. For example, previous research such as Raddick et al. 2009, Haklay 2011 and Ridge 2013 can be built upon by closer attention to our contributors' understanding of the subjects they engage in through our project.

The framework will be applied using a mixed quantitative and qualitative approach, and by combining more 'traditional' methods borrowed from the social sciences with "natively digital" ones (Rogers 2013) that (with due attention to ethical considerations) harvest cultural tastes and practices from social networks and try to understand how they influence community formation processes.

References

- Bonney, R., H. Ballard, R. Jordan, E. McCallie, T. Phillips, J. Shirk, C.C. Wilderman.** (2009). *Public participation in scientific research: Defining the field and assessing its potential for informal science education*. A CAISE Inquiry Group Report. Washington: Center for Advancement of Informal Science Education (CAISE).
- Haklay, M.** (2011). *Classification of Citizen Science activities* Available at: povesham.wordpress.com/2011/07/20/classificationofcitizenscienceactivities (accessed 31 October 2013).
- Mansell, R.** (2012). *Promoting access to digital knowledge resources: managing in the commons*. International journal of the commons, online. ISSN 18750281 (In Press).
- Raddick, M., Jordan, G. Bracey, K. Carney, G. Gyuk, K. Borne, J. Wallin, S. Jacoby.** (2009). *Citizen science: Status and research directions for the coming decade*. In Astro2010: The Astronomy and Astrophysics Decadal Survey, Vol. 2010.
- Ridge, M.** (2013). *From Tagging to Theorizing: Deepening Engagement with Cultural Heritage through Crowdsourcing*. Curator 56 (4): 435450.
- Rogers, R.** (2013). *Digital Methods*. Cambridge, MA: MIT Press.
- <http://micropasts.org>
<http://github.com/findsorguk>
 Built on the Pybossa platform – dev.pybossa.com
crowdsourced.micropasts.org
discourse.org
neighbor.ly
catarse.me/en

Treasure Challenge: an archaeological video conferencing journey

Pett, Daniel Edward John
 The British Museum, United Kingdom

Kelland, Katharine Louise
kkelland@britishmuseum.org
 The British Museum, United Kingdom

In 2014, the British Museum's department of Learning, Volunteers and Audiences will launch an innovative video conferencing activity: "*Roman Britain Treasure Challenge*" developed in collaboration with the Portable Antiquities Scheme (PAS) and the Department of Britain, Europe and Prehistory. This new activity will bring archaeology to life within primary schools in the United Kingdom from the British Museum's Samsung Digital Discovery Centre (SDDC)³, a dedicated facility for families and school children to utilise digital technology to enrich their visit. This collaboration between departments is the first of its kind for a digital educational activity in such a venerable institution and could open new avenues for researchers, curators and education practitioners. The activity will reach out to schools that may not have visited or be able to visit the actual site of the British Museum, thus facilitating their participation and broadening scope for interaction with audiences.

Building on the digital experience gathered through the delivery of ICT and through the use of the PAS's innovative and award winning website; curatorial and scientific staff and the world leading SDDC's museum educational programmes, a new activity has been formulated: Roman Britain Treasure Challenge. This schools session will be aimed at teaching Key Stage 2 children (between the ages of 7 to 11) a variety of life and ethical skills based around the amazing discovery of the Frome Hoard of 52,503 late Roman coins (Bland, Booth & Moorhead 2010) and the working of the legal processes of the Treasure Act (DCMS 1996). The Frome hoard has been acquired by Somerset Museum, and it is now partially on display in their new galleries, whilst the remainder of the hoard is still housed within the British Museum. The department of Conservation and Science's team of conservators are working on stabilising, cleaning and providing preventative care for many of the coins, before they are returned to Somerset for display.

The Portable Antiquities Scheme has been recording small finds of archaeological discoveries in England and Wales since 1997 and its database (accessed online) has records for over 930,000 objects discovered by the general public whilst pursuing their hobbies (for example gardening, walking or metal-detector) or going about their daily life. These Open Data provided through this system is now providing the basis for a wide variety of research (over 380 projects are now using these data) and innovative visualizations have been produced; for example see the recently released Lost Change application. The close ties between the British Museum's departments have made it possible to collaborate on this activity. Images, raw data

and video footage are combined within a predefined framework to create a structured learning environment.

The *Roman Britain Treasure Challenge session* begins with a discussion based around the question 'What do you think of when you hear the word treasure?' Typically this leads to descriptions that, for example include words such as 'gold' and 'how much is it worth?' (A phrase that is very alien to an archaeologist!) The session aims to dispel this perception, to challenge pre-conceived ideas of Treasure and impress on the young, what is archaeologically important. The session then proceeds through a very brief introduction to the fact that there is a Treasure Act (created in 1996 and replacing the old law of Treasure Trove), a legal definition of Treasure (note the capital T as Treasure is a concept) and a strict procedure that is followed to determine if a find is Treasure. In 2012, 969 cases of Treasure were reported to the British Museum's Treasure Registrar and 26 in Wales (DCMS 2013); therefore this process is one that could conceivably be experienced by session participants in the future, if they were lucky enough to discover Treasure.

Within the session, children are given simple, but structured tasks based around the discovery process; for example choosing who will comprise the excavation team, formulating and delivering a security strategy and researching the coins to determine which period the hoard is likely to be from (using replica coins posted to the school in advance). Taking the example of choosing the excavation team further, this is part of a wider activity to choose three teams; the excavation and discovery team, protecting the artefacts team and a research and display team. The list of possible candidates include people such as the finder, an archaeologist, the PAS Finds Liaison Officer, an illustrator, conservators and the museum director. Using an information sheet and knowledge gleaned from previous discussions the children (in their own class teams) must make the decision of who to allocate to each team. Who is vital for each stage of the artefacts journey? Who must not be left behind? It is dependent on the children to use knowledge gained during the session, discussion within the group and decision making skills to ensure that best people are tasked with the care of the Frome Hoard.

The session also includes a hands-on element as part of its interactivity. The school teacher is given a choice between two activities based around the study of coins (or numismatics); they can either decipher and interpret the inscription on a paper or digital representation of a coin from the hoard or replica handling coins are sent to the school and the children have to examine them and attempt to place them on to a timeline in either date or issuer (usually an Emperor) order.

The session uses a variety of ICT equipment including a dedicated equipment trolley which has been configured with new equipment and set up uniquely for this session with the video conferencing equipment and other equipment such as a 'visualiser' for close up viewing of real Roman coins from the Museum's handling collection (provenance is known for these.). The participating schools will use their own ICT equipment in either their ICT suites or within classrooms. This leads to a fully-fledged activity which will bring the archaeological process to life using presentations, audio-visuals (such as pre-recorded video using trained actors and archaeologists), replica and real coins and a trained museum educator to deliver the session. Using images that the PAS has disseminated via Flickr and also through its own database, children can take home information about museum and archaeological content to display in their school, or at home and easily query our resources in external contexts (for example one can search via post code or upon their local environs).

This paper will discuss how this activity was developed over the previous year (drawing inspiration from the National Space Centre's educational programmes¹, and further afield) with specific reference to teaching methods, curriculum suitability, equipment selection (in conjunction with the SDDC's commercial contracts that ties us to using Samsung manufactured and branded equipment), archaeological practice and the legal aspects of the Treasure Act (1996). It will show how some intrinsically complex situations are broken down for the younger audience, how some practical choices had to be made to enable the activity to be executed. It will also discuss

whether this initiative could be seen as a success, whether it could be replicated in other museums or archaeological facilities, whether it could be delivered to different audiences (for example with community archaeology groups or university students) and approximately how much staff time and budget was expended on development (some costs cannot be calculated as services are provided as 'in-kind').

This paper will also discuss how this type of session is evaluated and subsequently improved via feedback from participants (children and teachers) and observation from British Museum staff and Portable Antiquities Scheme employees. It will also show how this improvement process impacts on both staff and delivery time in a hectic term time schedule (planned months in advance.) This type of activity is a new venture for the British Museum, at the time of writing this abstract, it is not known how successful this session will have been; presenting the results will be a challenge to the authors and the proposed content will be subject to change.



References

Bland, R., Booth, A. & Moorhead T.S.N.M. (2010) "The Frome Hoard" London: British Museum Press

Department for Culture, Media & Sport (DCMS), (1996) "The Treasure Act 1996 Code of Practice (2nd Revision) England and Wales" London: HMSO

Department for Culture, Media & Sport (DCMS), (2013) "Reported Treasure Finds 2011 & 2012 Statistical Release" London: HMSO

http://www.britishmuseum.org/whats_on/events_calendar/event_detail.aspx?eventId=994&title=Roman%20Britain%20treasure%20challenge&eventType=School%20digital%20session

<http://finds.org.uk>

http://www.britishmuseum.org/learning/samsung_centre.aspx

<http://tracemedia.co.uk/lostchange>

<http://flickr.com/finds>

<http://finds.org.uk/database>

<http://education.spacecentre.co.uk/virtualclassroom/video-conferencing>

On Metaphor in Text Visualization Prototypes

Peña, ,Ernesto

ernesto.pena@gmail.com

University of British Columbia, Canada

Brown, Monica

mm2brown@gmail.com

University of British Columbia, Canada

Dobson, Teresa

teresa.dobson@ubc.ca

University of British Columbia, Canada

The content of this presentation develops from previous research and writing on the "Glass Cast," a prototype for mapping and visualizing complex knowledge networks in which time is crucial, and on "PlotVis," an interactive visualization resource for displaying XML-encoded fictional narratives that departs from ways of modeling narrative typical of literary pedagogy (see, e.g., Dobson, Michura, Ruecker, Brown, and Rodriguez 2011). Digital humanities scholars (e.g. Zepel 2013) have argued that by thinking about visualization as a tool, we can gain insight into visualization's purpose in DH scholarship, particularly as a form of "visual thinking." This perspective on visualization informs the approach we adopt in this presentation. Specifically, we describe aspects of the development of two visualization prototypes, offering a detailed overview of how these prototypes work. We also theorize visualization in terms of the concept of metaphor, examining the implications for visualization of the metaphors implicit in the respective structures of our experimental prototypes. How does metaphor shape and perhaps even constrain perceptions of the purpose of the two visualization prototypes we discuss? In attending to questions of metaphor and visualization, we contribute to ongoing theorization of the role of visualization, broadly conceived, in digital humanities scholarship,

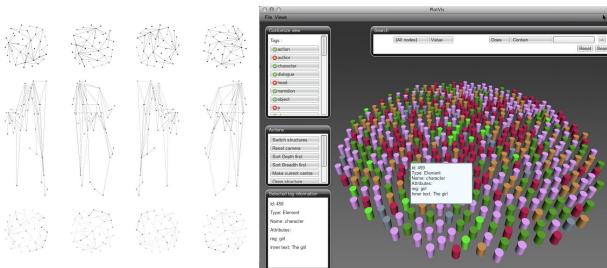


Fig. 1: Rendering from 3D Model, "Glass Cast" Prototype (Left); Still Image of XMLEncoded Literary Narrative from "PlotVis" Prototype (Right).

Elsewhere (Peña and Dobson 2013), we elaborate on the conceptualization of the Glass Cast as part of an effort to trace the usage of "visual literacy" and related notions through time and across different knowledge domains. As we explain, the aim of this experimental prototype is to represent networks chronologically and to facilitate the examination and exploration of these networks from different views in a three-dimensional visualization environment (see Figure 1). The Glass Cast serves as a tool for tracking the origin and mobilization of concepts across time and through different disciplinary fields, a key challenge in humanities and social sciences scholarship. Enabling the examination of knowledge networks along the lines of time, discipline, and connection between texts and authors required a three-dimensional representation.

In order to facilitate such a display, we adopted the metaphor of the "glass cast," thinking in particular about those types of casts in which three-dimensional figures have been impressed in the core of geometric shapes such as cylinders and polyhedrons. The use of such a metaphor in the Glass Cast visualization is intended to structure user interactions around the kinds of experience typical of interactions with the artifact from which the interface take its name. Yet the metaphor of the glass cast is also meant to leverage the understanding of the affordances of the prototype for those who interact with it; the Glass Cast namely (but not exclusively) provides not only a general reading of the data that is showcased, but also permits the display of another level of information depending on the side of the prism that faces the reader. Metaphor is thus germane to any understanding of the purposes of the Glass Cast as a scholarly visualization tool.

The Glass Cast aims to showcase abstract relationships between data that are hardly (or impossibly) acquirable from non-visual means, whether because the nature of the data is different and therefore these relationships are not explicit, or otherwise. This property of some visual resources to communicate relationships and the "intellectual skill" required to read or interpret them is what Balchin and Coleman (1966) defined as graphicacy. Shaped through the chosen

metaphors, these abstract relationships reveal new layers of information and patterns from either new or already existing data. In the particular case of the Glass Cast, we see this metaphor as a tool to (amongst other possible outcomes) seize the possibilities that the new media gives us to look back and through disciplines in a single artifact. In this context, visualization provides a powerful means of displaying the complex processes through which new disciplines form and at the same time constitutes a rethinking of disciplinarity.

PlotVis, which we developed as part of an interdisciplinary project on reading, writing, and teaching complex fiction, emerged out similar concerns with disciplinarity and, in particular, with disciplinary practices. Specifically, the prototype was designed to augment conventional approaches to teaching the concept of plot in fiction, which have tended to rely on a static visualization known as Freytag's Pyramid. Unlike this static, linear representation in print media, PlotVis, a digital tool, allows users to model and interact with XML-encoded literary narratives in three dimensions. A number of different metaphors (e.g. the Fibonacci Series, tubes, building blocks, cells, the cabinet of curiosities) dictate the structures of the multiple views that can be used to visualize encoded text (see Figure 1).

Incorporating multiple metaphors into the PlotVis prototype to facilitate a range of views on a single encoded text (or corpus of encoded texts) underscores an important point, which is that the ways in which plot is visualized informs and sometimes constrains a reader's understanding of the very concept of plot. Ways of modeling narrative that have traditionally relied on the Cartesian graph, wherein time is plotted on the x axis and the fortunes of the hero or the disposition of the action (whether it is 'rising' or 'falling') is plotted on the y axis, have for example led many readers to expect literary narratives to conform to this sequence of events. Given this consequence of visualization, PlotVis was (like the Glass Cast) developed to radically alter and, importantly, expand user engagement in order to do justice to the complexity of plot as a central concept of literary criticism. This presentation opens up an opportunity to elaborate on the influences of metaphor on visualizations in PlotVis. While the metaphors serve as tools for displaying different views on the same text, how might PlotVis metaphors structure users' understanding of the goals of literary criticism?

In conclusion, in the case of both prototypes, we consider how visualization mediates participation within a scholarly discipline and ask, "What does metaphor have to do with this?" By participation, we mean the kinds of scholarly engagement a visualization interface enables and how such engagement might be informed by an interface's metaphor entailments (see, e.g., Kerr et al., 2013). This presentation thus provides a forum for exploring the relationship not only of metaphor to visualization, but also of visualization to disciplinarity.

References

- Balchin, W. G. V., & Coleman, A. M. (1966). *Graphicacy should be the fourth ace in the pack Cartographica: The International Journal for Geographic Information and Geovisualization*, 3(1), 23-28. doi: 10.3138/C7Q0-MM01-6161-731
- Dobson, T. M., Michura, P., Ruecker, S., Brown, M., and Rodriguez, O. (2011). *Interactive visualizations of plot in fiction*. Visible Language, 45: 169–91.
- Kerr, K., Hausman, B. L., Gad, S., & Javen, W. (2013). *Visualization and rhetoric: Key concern for utilizing big data in humanities research*.
- Peña, E., and T. Dobson. (2013). *Visualizing knowledge networks: The Glass Cast prototype*. Paper presented at Research Foundations for Understanding Books and Reading in the Digital Age. New York: September 26-27
- Zepel, T. (2013). *Visualization in the digital humanities: Tool or 'discipline'?* Retrieved from <http://www.hastac.org/blogs/tzepel/2013/01/09/visualization-digital-humanities-tool-or-%E2%80%98discipline%E2%80%99>

Modelling digital editing: of texts, documents and works

Pierazzo, Elena

elena.pierazzo@kcl.ac.uk
King's College London

Noël, Geoffroy

geoffroy.noel@kcl.ac.uk
King's College London

Editing is one of the most important activities we have to perform as digital humanists: with the entire literary and historical production facing remediation, the need for a theoretical as well as practical understanding of what it is at stake and what does it mean to create a digital (scholarly) edition is of crucial importance. Many contributions in the past have dealt with the issue of what a text is, what a document is, how they relate to each other and which are the implications of their ontological status with respect to the work they manifest¹ (2 3 4 5 6 7 8 9 10 11 12 to name a few). The topic has been tackled from different points of view: sociological, cultural, psycholinguistic, philosophical, historical and computational. Why, then, is it necessary to return to the same topic once again? Because while some if not most of the previous contributions have touched upon it, none has tried to account for the whole concept of digital editing, and if a model is "a representation of something for purposes of study, or a design for realizing something new" (p. 21)¹³, a new purpose of study will require a new model. In order to perform an activity with the help of a computer, in order to digitize a workflow, such an activity (i.e. editing) needs to be modelled, as there is "the fundamental dependence of any computing system on an explicit, delimited conception of the world or 'model' of it" (p. 210)¹⁴.

The only existing model of text that has been developed explicitly for computational purposes is represented by the so-called OHCO model¹⁵, the limitations of which are widely known^{16 17 18 19 20}. However, as it reflects upon the fundamental way of functioning of the most important technology used so far for the production of digital editions, i.e. XML, in spite of its inadequacy, it still represents a fundamental approach for editorial endeavours: if one edits using XML and TEI, then one will have to adopt some sort of model which strongly relates to the OHCO model. But besides all the issues already pointed out by earlier critics, there is a series of facts and entities which are outside the scope of the OHCO model but which are nevertheless of fundamental importance in scholarly editing; namely, what is an edition? What is a work? What is the relationship between an edition and the text it edits, and the work the text represents? What is the function of the reader and of the editor in establishing the text and the edition?

The edition of a text, any text, embodies a model of the work which the text represents. In the same way that a map can be considered a model of the earth built for a specific purpose, an edition of a text can be considered as a model of the text itself, because it represents a selection of the infinite features of a text according to particular point of view, scholarly or not. Selecting also means simplifying: a model is necessarily a simplification of a real life object, which makes it more apt to analysis and manipulation. In a contribution from 2009 Michael Sperberg-McQueen declared that there are three things to consider when we edit a text (p. 31)²¹:

1. There is an infinite set of facts related to the work being edited.
2. Any edition records a selection from the observable and the recoverable portions of this infinite set of facts.
3. Each edition provides some specific presentation of its selection.

In saying this he declared an edition to be a model of a work, where the act of selecting features from the uninterrupted continuity of the reality is the defining act of modelling, the purpose of which in turn is to provide a discrete selection of facts to be interpreted. The present research stems from this

consideration and proposes a new, comprehensive conceptual model of the editorial domain, which could be called the *Digital Editing Model*, or DEM.

The conceptual model deals with the following entities:

- **Documents**, i.e. physical objects that contain some sort of information; therefore a book is a document, as is a leaf with some writing on, a stone, and so on. More generally, a document is a physical object that has some text on it, or more formally, a Verbal Text Bearing Object, or VTBO. This definition willingly and knowingly omits non-verbal documents, as the object of the present research is to analyze and model written and verbal texts with the purpose of editing them.[1]
- **User-function**: any type of human interaction with the documents. The entity represents set of functions, more than human beings, such as, for instance, reading, editing, collecting, preserving, transcribing, analyzing etc.

As seen before, documents present an infinite set of **facts**. A user-function selects a subset of these facts and, according to an organizing principle, groups them into **dimensions**. As the dimensions that are potentially observable in a document are defined by the user-function's purpose, consequently it is impossible to draw a stable and complete list of such dimensions; however, for the purpose of exemplification, such a list could include linguistic, semantic, literary, genetic, iconographic, codicological, and palaeographical dimensions.

The interaction of user-functions with documents generates

- **Document models**: the meaning(s) that user-functions give to the subset of dimensions they derive from a document and that they consider interesting. If the subset of dimensions considered by the user includes the verbal content of the document, such a document's model is defined as a **text**.
- **Works**: an editorial statement of the fact that a number of documents aim to contain more or less the same verbal content. The sum of all possible texts derived from such documents, in a one-to-one or many-to-one relation, constitute the work.

The model does not necessarily need an **author-function**, but it may, if the user-editor postulates it. If present, the author-function performs two main sub-functions, one *in posse* and one *in esse*, where the latter represents the activities of producing some of the facts present in the documents, especially, but not exclusively, the ones concerning the verbal content of the documents. The function *in posse* concerns instead the authorial intention, namely what the author-function wanted to produce but did not, or, if it did, the evidence for this is lost. It is *in posse* because it is unachieved (or perhaps is achieved but with no way of knowing this).

The DEM model sketched out here, will be supplemented by a second one on text transmission (how text migrates from one support to the next), where emerging theories of transcription will also fit very well^{22 23 24 25 26}. However, these latter theories can also be thought as an instance of the user-function, encompassing the building of texts from the infinite set of facts available from the documents, and therefore are integrated into the DEM. In its full version, DEM will also deal with concepts such as versions and derivative works.

In conclusion the proposed DEM contributes to the elaboration of a set of models required by the renewed work and workflow of digital editing, in dialog with previous scholarly elaborations, providing a holistic and agnostic base for the understanding of the digital editorial endeavour.

[1] A similar but more generalized definition is given by Huitfeldt and Sperberg-McQueen, who prefer to speak of 'marks' on a document, rather than of 'verbal text': 'By a document we understand an individual object containing marks. A mark is a perceptible feature of a document (normally something visible, e.g. a line in ink)'²⁷.

References

1. DeRose, S. J., D. G. Durand, E. Mylonas, and A. H. Renear (1990). *What is Text, Really?* Journal of Computing in Higher Education 2:1 3-26.

2. **Shillingsburg, P. L.** (1991). *Text as matter, concept, and action*. Studies in Bibliography, 44:31–83.
3. **Shillingsburg, P.** (2006). *From Gutenberg to Google: Electronic Representations of Literary Texts*. Cambridge: Cambridge University Press.
4. **Caton, P.** (2013). *On the term text in digital humanities*. Literary and Linguistic Computing, 28(2):209–220.
5. **Gabler, H. W.** (2012). *Beyond author-centricity in scholarly editing*. Journal of Early Modern Studies, 1:15–35.
6. **Eggert, P.** (2009). *Securing the Past: Conservation in Art, Architecture and Literature*. Cambridge: Cambridge University Press.
7. **Tanselle, G. Thomas.** *A Rationale of Textual Criticism* Philadelphia: University of Pennsylvania Press, 1989. 104 pp.
8. **Sperberg-McQueen, C. M.** (2009). *How to teach your edition how to swim*. Literary and Linguistic Computing, 24(1):27–52.
9. **Robinson, P. M.** (2009). *What text really is not, and why editors have to learn to swim*. Literary and Linguistic Computing, 24(1):41–52.
10. **Robinson, P. M. W.** (2013). *Towards a theory of digital editions*. Variants 10. 105–132.
11. **Barthes, R.** (1968). *La Mort De l'Auteur*. Manteia (4e trimestre).
12. **Ong, J. Walter** (1975). *The Writer's Audience is Always a Fiction*. PMLA, Vol. 90/1, pp. 9–21
13. **McCarty, W.** (2005). *Humanities Computing*. Palgrave Macmillan.
14. **McCarty, W.** (2004). *Modeling: a Study in Words and Meanings*. in Companion to the DigitalHumanities. Blackwell.
15. **DeRose, S. J., D. G. Durand, E. Mylonas, and A. H. Renear** (1990). *What is Text, Really?* Journal of Computing in Higher Education 2:1 3–26.
16. **Huitfeldt, C.** (1994). *Multi-Dimensional Texts in a One-Dimensional Medium*. Computers and the Humanities, 28(4–5). Humanities Computing in Norway. 235–241.
17. **Pichler, A.** (1995). *Transcriptions, texts and interpretation*. In Johannessen, K. and Nor- denstam, T., editors, Culture and Value: Philosophy and the Cultural Sciences, pages 690–695. Austrian Ludwig Wittgenstein Society, Wien.
18. **Renear, A. H., Mylonas, E., and Durand, D.** (1996). *Refining our notion of what text really is: The problem of overlapping hierarchies*. In Ide, N. and Hockey, S., editors, Research in Humanities Computing. Oxford University Press.
19. **Pierazzo, E. and Stokes, P. A.** (2010). *Putting the text back into context: a codicological approach to manuscript transcription*. In Fischer, F., Fritz, C., and Voelger, G., editors, Kodikologie und Palographie im Digitalen Zeitalter 2 - Codicology and palaeography in the digital age 2, pages 397–430. Books on Demand, Norderstedt.
20. **Deegan, M. and Sutherland, K.** (2009). *Transferred Illusions*. Digital Technology and the Forms of Print. Ashgate, Farnham.
21. **Sperberg-McQueen, C. M.** (2009). *How to teach your edition how to swim*. Literary and Linguistic Computing, 24(1):27–52.
22. **Huijfeldt, C., and C. M. Sperberg-McQueen** (2008). *What is transcription?* Literary and Linguistic Computing. 23 (3). 295–310. doi:10.1093/linc/fqn013
23. **Huitfeldt, Claus, Yves Marcoux and C. M. Sperberg-McQueen** (2010). *Extension of the type/token distinction to document structure*. In Balisage: The Markup Conference 2010. held August 3–6, 2010 in Montréal, Canada. In Proceedings of Balisage: The Markup C
24. **Sperberg-McQueen, C. M.. Claus Huitfeldt, and Yves Marcoux** (2009). *What is transcription? Part 2*. Talk given at Digital Humanities 2009, College Park, Maryland. Slides on the Web at blackmesatech.com/2009/06/dh2009/.
25. **Caton, Paul** (2013-. *Pure transcriptional encoding*. Paper given at Digital Humanities 2013, Lincoln, Nebraska.
26. **Caton, P.** (2013). *On the term text in digital humanities*. Literary and Linguistic Computing, 28(2):209–220.
27. **Huitfeldt, Claus, and C. M. Sperberg-McQueen**. (2008) *What is transcription?* Literary & Linguistic Computing 23.3: 295–310.

Cultural text mining: using text mining to map the emergence of transnational reference cultures in public media repositories

Pieters, Toine

t.pieters@uu.nl

Utrecht University, The Netherlands

Verheul, Jaap

j.verheul@uu.nl

Utrecht University, The Netherlands

Introduction

This paper discusses the research project Translantis, which uses innovative technologies for cultural text mining to analyze large repositories of digitized public media, such as newspapers and journals.¹ The Translantis research team uses and develops the text mining tool Texcavator, which is based on the scalable open source text analysis service xTAS (developed by the Intelligent Systems Lab Amsterdam). The text analysis service xTAS has been used successfully in computational humanities projects such as Political Mashup, WAHSP, BILAND, and DutchSemCor. Within the context of the Translantis project, xTAS, coupled to Elasticsearch, will be further developed. Future versions will include clustering concepts and sentiment mining of issues in public debates. Translantis researchers are using Texcavator to detect and track cultural references in large textual corpora.

Use case: mining transnational references in public discourse

In order to test the potential of cultural text mining, Texcavator will be used to analyze the role of reference cultures in debates about social issues and collective identities. The central use case of this project is the emergence of the United States in public discourse in the Netherlands from the end of the nineteenth century to the end of the Cold War. This concept of reference culture is be used to discuss long-term asymmetrical processes of cultural exchange involving dimensions of power and hegemony. The concept recognizes the fact that some cultures assume a dominant role in the international circulation of knowledge and practices, offering or imposing a model that others imitate, adapt, or resist.

Reference cultures are mental constructs that do not necessarily represent a geopolitical reality with an internal hierarchy and recognizable borders. These culturally conditioned images of trans-national models are typically established and negotiated in public discourses over a long period of time. However, the specific historical dynamics of reference cultures have never been systematically analyzed and hence are not fully understood. To explore these dynamics, this project asks three interrelated questions.

1. How can e-tools be used to map trends and changes in relation to the economic power, cultural acceptance, and scientific and technological impact of the United States as reference culture?
2. How does public discourse reflect and influence the emergence and impact of reference cultures?
3. How were ideas, products and practices associated with the United States valued in Dutch public discourse between 1890 and 1990?

We propose that the key to understanding the emergence and dominance of reference cultures is to chart the public discourse in which these collective frames of reference are established. Text mining methodologies allow us to trace changes in “big data” repositories of public media, such as newspapers, journals, and other periodicals. Central to this project is the large digital data collection of the National Library

of the Netherlands (KB), which contains 9 million newspaper pages and over 1.5 million journal pages². This large collection of serialized historical texts, which have been OCR-ed and provided with meta-tags, allows us for the first time to study long-term developments and transformations in national discourses in a systematic, longitudinal, and quantifiable way, by using innovative text-mining tools.

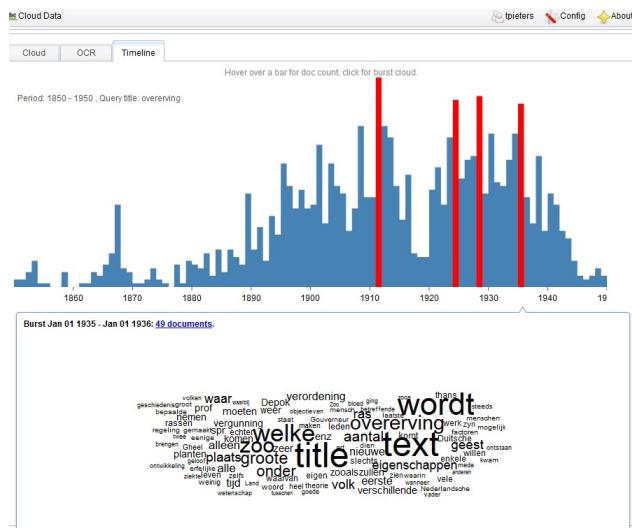
Methodological innovations and challenges

The semantic text mining tool Texcavator has direct access to historical textual repositories and is able to handle queries on-the-fly, and to produce visualization such as timelines and word clouds based on integrated topic modeling and NER modules. This allow us to test the value of qualitative heuristic models and to pair them in a meaningful fashion with quantitative methodology. Some of the methodological challenges involve the calibration between close and distant reading, the normalization of search results from unevenly distributed historical media, and adjusting for lexicological changes that affect the accuracy of sentiment mining and concept mining.

First results indicate the ability to mine “hidden debates” in public media in a bottom-up (inductive) manner, based on the footprints that used terms leave behind. More importantly, the tool is innovative in that it pinpoints continuities and discontinuities in public discourse, for instance by showing variations in the context in which key terms are used, and changes in sentiment values of words over time. We argue that this marks a promising transition from text mining to “concept mining” and new forms of cultural text mining that go beyond already established mining features.

Conclusions

We will demonstrate that semantic mining of big data open new vistas in historical research because they (a) provide a robust framework for producing new vistas on macro history; and (b) can be complemented with numerical data sets provided by other researchers, for example on economic and social trends. This, ultimately, is the transformative promise of digital humanities as a multi-dimensional window on political, economic, and social change.



References

1. **Translantis: Digital Humanities Approaches to Reference Cultures; The Emergence of the United States in Public Discourse in the Netherlands, 1890-1990** (funded by the Netherlands Organization for Scientific Research), www.translantis.nl.
 2. kranten.delpher.nl/

Aiden, Erez, and Jean-Baptiste Michel (2013). *Uncharted: Big Data as Lens on Human Culture*. New York: Penguin.

Balog, K., M. Bron and M. de Rijke (2011). ‘Query Modeling for Entity Search Based on Terms, Categories and Examples,’ *ACM Transactions on Information Systems* 29, no. 4, Article 22, November 2011.

Dougherty, M., E.T.Meyer, C. Madsen, C. van den Heuvel, A. Thomas, and S. Wyatt (2010). *Researcher Engagement with Web Archives: State of the Art*. London: JISC.

Eijnatten, Joris van, Toine Pieters, and Jaap Verheul (2013). “*Big Data for Global History: The Transformative Promise of Digital Humanities*.” *Low Countries Historical Review/BMGN* 128-4: 55-77.

Hernandez, J.F., A.K. Mantel-Teeuwisse, G.J.M.W. van Thiel, S.V. Belitser, J.A.M. Raaijmakers and T. Pieters (2011). *Publication trends in newspapers and scientific journals for SSRIs and suicidality: a systematic longitudinal study*. *BMJ OPEN* 6, no. 1.

Huurnink, B., L. Hollink, W. van den Heuvel and M. de Rijke. ‘Search Behavior of Media Professionals at an Audiovisual Archive: A Transaction Log Analysis.’ *Journal of the American Society for Information Science and Technology* 61, no. 6 (June 2010): 1180-1197.

Huijnen P., F. Laan, M. de Rijke, and T. Pieters. *A digital humanities approach in the history of science; eugenics revisited in hidden debates by means of semantic text mining*. Histoinformatics, Springer (forthcoming, 2013).

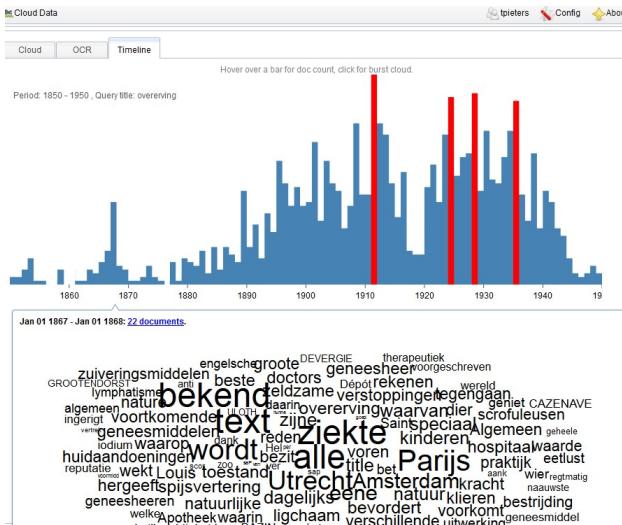
Jijkoun, V., M. de Rijke and W. Weerkamp. ‘Generating Focused Topic-specific Sentiment Lexicons,’ 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010), July 2010.

Meij, E., M. Bron, L. Hollink, B. Huurnink and M. de Rijke. “*Mapping queries to the Linking Open Data cloud: A case study using DBpedia*.” *Journal of Web Semantics* 9, no. 4 (November 2011): 418-433.

Pieters, T., and S. Snelders. “*Standardizing psychotropic drugs and drug practices in the twentieth century: Paradox of order and disorder*.” (2011) *Studies in the History and Philosophy of the Biological and Biomedical Sciences* 42: 412-415.

Snelders, S., and T. Pieters. (2011) “*Speed in the Third Reich: Metamphetamine (Pervitin) Uses and a Drug History From Below*.” *Social History of Medicine*. First published online: February 19.

Thomas, A., E.T. Meyer, M. Dougherty, C. van den Heuvel, C. Madsen, and S. Wyatt (2010). *Researcher*



Engagement with Web Archives: Challenges and Opportunities for Investment. London: JISC

Verheul, Jaap (2010). "Through Foreign Eyes." In *Discovering the Dutch: On Culture and Society of the Netherlands*, edited by Emmeline Besamusca and Jaap Verheul, 267-77. Amsterdam: Amsterdam University Press.

Aplicación del análisis dinámico de redes científicas al estudio de la evolución de la investigación española relacionada con el descriptor "historia del arte" durante 1976-2012, según ISOC.

Pino-Díaz, José

Dpto. Historia Arte. Universidad de Málaga, Spain

Cruces-Rodríguez, Antonio

antonio.cruces@uma.es

Dpto. Historia Arte. Universidad de Málaga, Spain

Rodríguez-Ortega, Nuria

nro@uma.es

Dpto. Historia Arte. Universidad de Málaga, Spain

Bailón-Moreno, Rafael

Dpto. Ingeniería Química. Universidad de Granada. Spain.

1. Objetivo

El objetivo de la presente comunicación ha sido detectar las dinámicas de la investigación española relacionadas con la historia del arte en un rango cronológico que se extiende desde 1976 hasta 2012. Para ello, se ha utilizado como estrategia el análisis dinámico de redes científicas, lo que nos ha permitido extraer de uno de los corpus más representativos de la investigación histórico-artística en España (ISOC) un conocimiento implícito sobre la evolución y desarrollo de las estrategias temáticas y líneas de investigación relacionadas con la historia del arte como descriptor.

La elección de este rango cronológico no es arbitraria, pues la segunda mitad de la década de los setenta se ha conceptualizado tradicionalmente como un punto de inflexión en los estudios histórico-artísticos en España, al incorporarse metodologías y modelos interpretativos foráneos una vez iniciado el periodo aperturista tras el fin de la Dictadura de Franco, la llegada de todo un conjunto de traducciones de obras que desarrollaban modelos interpretativos distintos a los anteriormente establecidos, y la consolidación de los estudios de Historia del Arte en España como campo académico autónomo.

2. Marco teórico-metodológico

Una red científica está formada según (Latour, 1983) por el conjunto de actores de la red (investigadores, centros, revistas, temas de investigación, etc.) y por el conjunto de asociaciones establecidas entre ellos. Los actores independientemente de su naturaleza pueden definirse siempre mediante palabras. La red de palabras y de asociaciones establecidas entre ellas es manifestación del comportamiento de la Sociedad y de la estructura del Conocimiento. Los actores son, igualmente, entidades dinámicas que continuamente redefinen sus relaciones y, en consecuencia, la red socio-cognitiva que conforman. Con el devenir, los actores y las relaciones cambian y dan lugar a nuevas redes, y así se suceden unas a otras a lo largo del tiempo.

Latour define la Teoría Actor-Red, teoría sociológica sobre la generación de conocimiento científico, como "Sociología

de las Relaciones" (Latour, 2005). La teoría de la traducción (entendida esta como conversión, transformación, variación o cambio) estudia los cambios que se producen en las relaciones entre los actores de la red. Estos cambios en las relaciones entre los actores producen su aparición, fortalecimiento, equilibrio, debilitamiento o desaparición. La aparición de nuevos actores se produce por emergencia o por convergencia; el fortalecimiento se produce por convergencia o por evolución incremental; el equilibrio, por evolución estable; el debilitamiento, por evolución decremental o por divergencia; y la desaparición, por bifurcación o por exitus. A consecuencia de todos esos cambios, las redes tecnocientíficas de conocimiento se encuentran en continuo cambio (Ruiz-Baños, 1999).

Las relaciones naturales de coocurrencia de palabras en los textos científicos y técnicos constituyen una red tecnocientífica que puede ser analizada y cartografiada. El análisis y la visualización de estas redes es posible por el desarrollo de sistemas expertos denominados sistemas de conocimiento.

3. Criterios y proceso de análisis

El corpus documental de análisis lo constituyen los artículos de ISOC, <http://bddoc.csic.es:8080/isoc.html>, obtenidos mediante la búsqueda "historia del arte" en los campos título, resumen y descriptores, en el periodo 1976-2012. Este corpus está formado por 873 registros.

Se ha realizado el análisis dinámico de la red científica mediante el sistema de conocimiento Techné Coword; sistema de conocimiento que tiene su antecedente en Copaird (Bailón-Moreno, 2003), que a su vez tiene como precursor Leximappe (Law, Bauin, Courtial, & Whittaker, 1988); (Law & Whittaker, 1992). Para realizar el análisis de palabras asociadas, se ha creado un único campo formado por los descriptores, autores y revista de cada registro. Se han fijado los siguientes parámetros de análisis: ocurrencia mínima, 2; coocurrencia mínima, 2; tamaño mínimo de la subred, 2; y tamaño máximo, 12. Los subperiodos de estudio han sido: 1976-1983; 1984-1989; 1990-1995; 1996-2001; 2002-2007; y, 2008-2013.

Se han empleado técnicas KDD (knowledge databases discovery) y de text mining (coword analysis o análisis de palabras asociadas) para obtener las subredes de investigación más importantes y relevantes de cada subperiodo. El diagrama dinámico de subredes permite el análisis y la visualización de la evolución de los resultados de la investigación.

4. Resultados e interpretaciones

La gráfica de producción acumulada de documentos por año permite distinguir dos períodos de "inicio-expansión-agotamiento" sucesivos: uno de 1976 hasta 1986, y otro de 1987 hasta la actualidad (ver Figura 1)

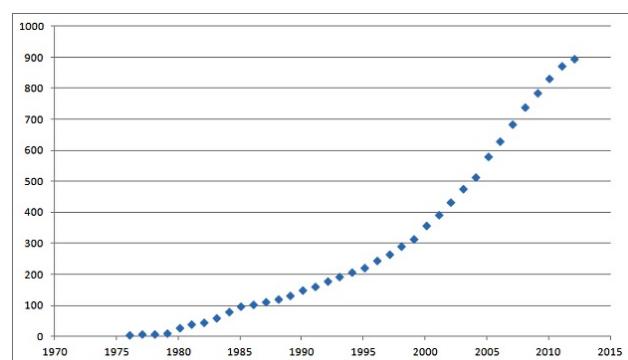


Fig. 1: Diagrama de producción anual acumulada de los documentos que incluyen "historia del arte" en su título, resumen o palabras clave.

Estos períodos son una constatación empírica de la Segunda Ley de Price o de Crecimiento Logístico de la Ciencia (de Solla Price, 1963), así como del concepto de cambio de paradigma, según la obra de Thomas Kuhn La Estructura de las Revoluciones Científicas (Kuhn, 1962). Esto último

se reafirma aun más si cabe por los profundos cambios conceptuales que ha sufrido la investigación relacionada con el descriptor "historia del arte" según el análisis de palabras asociadas que sido realizado para el presente estudio. Este análisis de palabras asociadas ha identificado los actores temáticos, temas o líneas de investigación en cada periodo, y las ha relacionado entre ellos dando lugar a diferentes salidas gráficas muy significativas. Los temas de investigación se muestran como redes de palabras que los han definido conceptualmente (ver Figura 2).

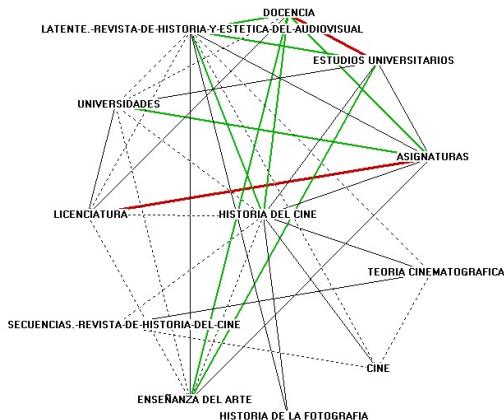


Fig. 2: Tema "historia del cine", tema motor de investigación del último periodo, 2008-2013

Los diagramas estratégicos han revelado las relaciones intertemáticas y su posición estratégica dentro del conjunto de la red (posición motora del campo científico, cuadrante superior derecho; posición transversal, cuadrante inferior izquierdo; posición accesoria, cuadrante superior izquierdo; y posición naciente o en decadencia, cuadrante inferior izquierdo) (ver Figuras 3, 4, 5, 6, 7 y 8). En los diagramas estratégicos se representan los temas de investigación; estos aparecen nombrados por el descriptor central de la subred que conforman y situados según los parámetros de centralidad y densidad de la misma.

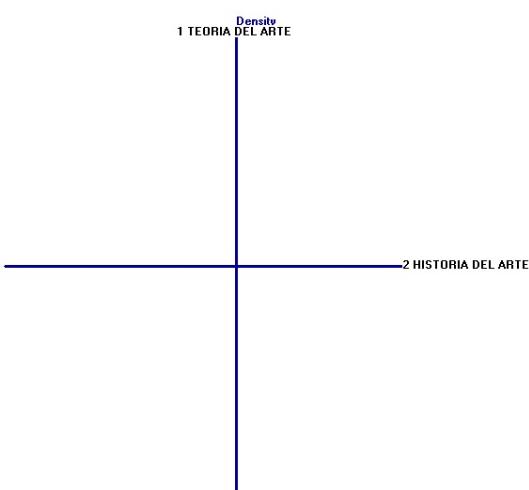


Fig. 3: Diagrama estratégico del periodo 1, 1976-1983.

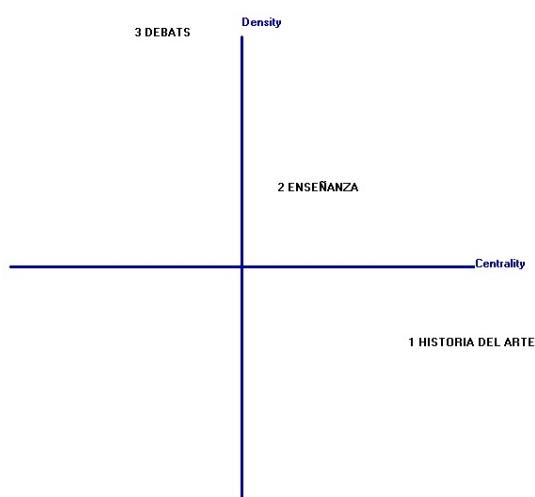


Fig. 4: Diagrama estratégico del periodo 2, 1984-1989.

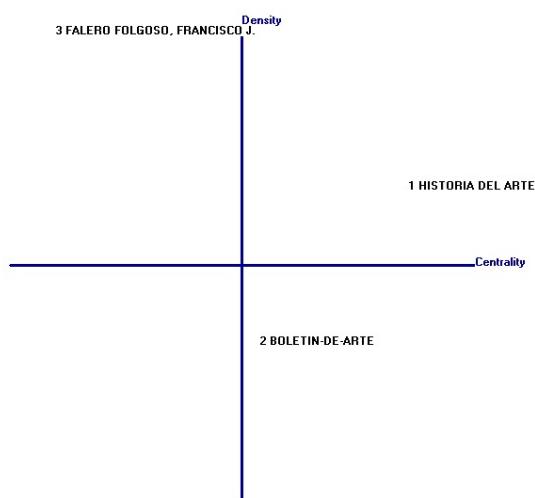


Fig. 5: Diagrama estratégico del periodo 3, 1990-1995.

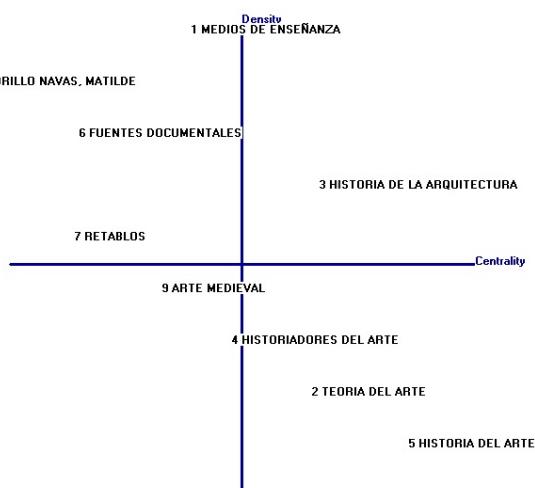


Fig. 6: Diagrama estratégico del periodo 4, 1996-2001.

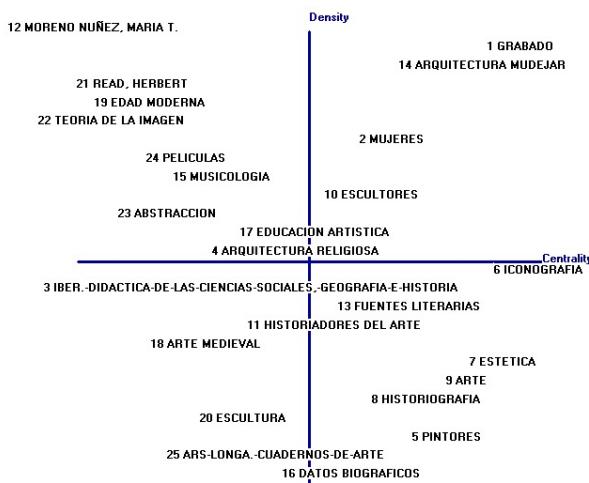


Fig. 7: Diagrama estratégico del periodo 5, 2002-2007.

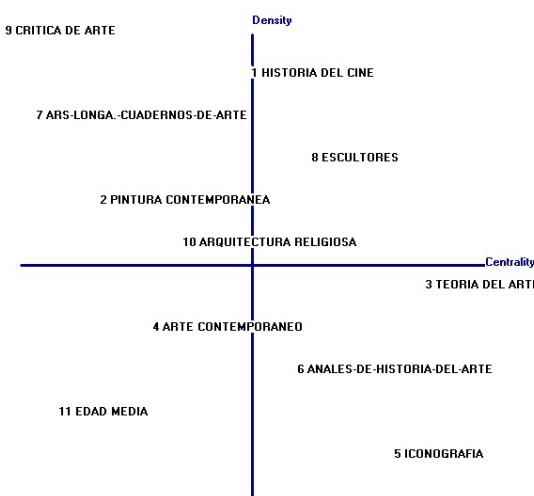


Fig. 8: Diagrama estratégico del periodo 6, 2008-2013.

Igualmente, el diagrama dinámico de las series temáticas ha evidenciado la evolución o traducción a lo largo del tiempo (ver Figura 9)

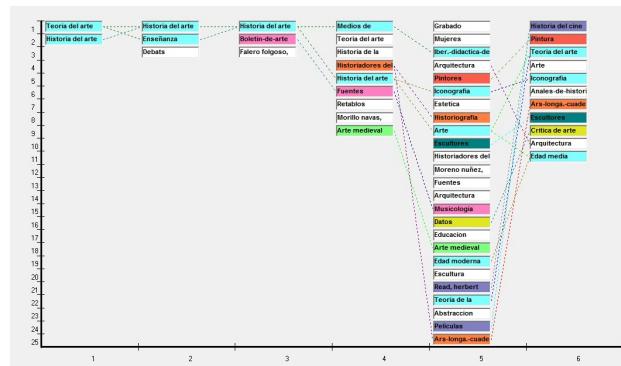


Fig. 9: Diagrama dinámico o de series temáticas. Los periodos son: 1, 1976-1983; 2, 1984-1989; 3, 1990-1995; 4, 1996-2001; 5, 2002-2007; y, 6, 2008-2013

Varias razones de tipo socioeconómico explican la existencia de los períodos anteriormente identificados y los cambios conceptuales y estratégicos sufridos por los actores temáticos en ellos: el dirigismo institucional de la investigación española que está sometido a los vaivenes de la economía general del país en cada momento, la incorporación española a la Unión Europea y a sus planes de apoyo a la investigación, la adopción de un sistema de evaluación de la actividad de los investigadores que impulsa la producción de artículos de

revistas, la redefinición de antiguas líneas de investigación y la emergencia de otras nuevas.

References

- Bailón-Moreno, R. (2003). *Ingeniería del conocimiento y vigilancia tecnológica aplicada a la investigación en el campo de los tensioactivos. Desarrollo de un modelo ciencimétrico unificado*. Granada: Tesis doctoral. Universidad de Granada.
- de Solla Price, D. J. (1963). *Little Science, Big Science*. New York: Columbia University Press.
- Kuhn, T. (1962). *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.
- Latour, B. (1983). *Give me a Laboratory and I Will Raise the World*. En K. Knorr-Cetina, & M. Mulkay, *Science observed: Perspectives on the Social Study of Science*. Londres: Sage.
- Latour, B. (2005). *Reassembling the Social. An introduction to Actor-Network Theory*. Oxford: Oxford University Press.
- Law, J., & Whittaker, J. (1992). *Mapping acidification research: a test of the co-word method*. *Scientometrics*, 23(3), 417-461.
- Law, J., Bauin, S., Courtial, J. P., & Whittaker, J. (1988). *Policy and the mapping of scientific change: a co-word analysis of research into environmental acidification*. *Scientometrics*, 14(3-4), 251-264.
- Pino-Díaz, J., Jiménez-Contreras, E., Ruíz-Baños, R., & Bailón-Moreno, R. (Julio-Septiembre de 2011). *Evaluación de redes tecnocientíficas: la red española sobre Áreas Protegidas, según la Web of Science*. *Revista Española de Documentación Científica*, 34(3), 301-333.
- Pino-Díaz, J., Jiménez-Contreras, E., Ruíz-Baños, R., & Bailón-Moreno, R. (April de 2012). *Strategic knowledge maps of the techno-scientific network (SK Maps)*. *Journal of the American Society for Information Science and Technology*, 63(4), 796-804.
- Ruiz-Baños, R. (1999). *Las traducciones dinámicas de las series temáticas, propuesta de una clasificación*. La representación y la organización del conocimiento en sus distintas perspectivas: su influencia en la recuperación de la información. *Actas del IV Congreso ISKO*. Granada.

Incommensurability? Authorship, Style, and the Need for Theory

Plasek, Aaron

alp445@nyu.edu
New York University, US

1. The need for theory

Harold Somers and Fiona Tweedie pose the following question: If a vocabulary-based authorial attribution technique fails to attribute an original text and a pastiche to different authors, "is this because the pastiche is good, or because the technique is faulty" (Somers and Tweedie, 2003)? The question places computational techniques and literary concerns into a direct relationship, inviting a formulation that might "leap from [word] frequencies to meanings" (Craig, 1999), and so seems an ideal opportunity to explore and interrogate our assumptions surrounding literary interpretation, literary style, and genre. Moreover, a failure or refusal to attend to the question would seem to be a catastrophe for attribution studies: without a generalized set of criteria to critique how stylistic concerns influence an algorithm's effectiveness at identifying an author's statistical "fingerprints," the general validity of authorial attribution techniques will remain contested despite persuasive examples of authorial attribution techniques like Hoover (2003) and Garcia and Martin (2007). While scholars such as Jockers (2013) are interested in exploring how attributes such as lexical variability, word frequencies, word choice, and other statistical measures can be used as indicators of authorship, style,

genre, gender, and even nationality, there remains a paucity of theory to explain why these and other indicators happen to be more or less effective with respect to the finite sets of authors, books, and/or genres they are applied to. Somers and Tweedie approach their question pragmatically—that is, they subject Alice in Wonderland, Gilbert Adair's pastiche entitled Alice through the Needle's Eye, and several "con-trol" texts to a battery of authorship attribution techniques and report the results of their tests. They do not, however, provide a theoretical framework to understand why one technique would be more or less effective than another. Nor are their results generalizable to other cases because there does not exist a larger theoretical framework to understand how Somers and Tweedie's experiments may relate to a different set of originals and pastiches we might examine. The success or lack of success of the statistical techniques used to distinguish authorship has much to do with the idiosyncrasies of the individual texts and authors being considered and is frequently aided by our historical knowledge of existing texts for which author- ship is already known. But far too little effort has been devoted to developing a theoretical model that might provide us with a compendium of the possible ways our statistical methods might fail.

2. The interdependence of authorship and style

Somers and Tweedie's question highlights the contextual dependencies of our terms and the basic differences in assumption between nontraditional authorship attribution and computational stylistics. Effectiveness, for example, appears sensible in the context of authorship attribution techniques, but less so in the context of stylistics. Nontraditional authorship attribution techniques exist in a system for which the value of the question hinges on its falsifiability. Is a text of unknown authorship written by author X or Y given a set of existing texts written by both authors? There can only be one correct historical answer—which is usually only one author—and this correct answer is mutually exclusive to any other answer. Such "facts" are independent of the method we use to discover them. Alternatively, the question of pastiche quality—of whether the imitation is well or poorly done—is a question for which stylistics should provide an answer; nontraditional authorship attribution may also influence judgments of quality. When performed under the banner of literary studies, computational stylistics is concerned with interesting answers that point us to new interpretive insights about a particular text we are studying. These facts are less stable in that they depend on the methods which allow for their discovery and are not necessarily mutually exclusive.

Yet the epistemological distinctions above are countermanded by cases in which the concerns of authorship interpenetrate the concerns of style in ways that are difficult to generalize. When Erasmus declares a certain letter to be incorrectly at- tributed to St. Jerome based on the belief that "Jerome has a special quality about him, a kind of mental savour and temperament, a quality which may be felt rather than explained," and, earlier, when we see this "never-failing quality, his lively humour...which the learned admire in Cicero," the stylistic concern of quality is being used to determine authorship (1992, 80; see also Love 2002:18-22). Yet are issues of authorship and style always interrelated? The answer, I contend, is yes; in limiting cases where we have appeared to isolate these concerns it is because we have already (intentionally or unintentionally) picked our texts in such a way that separation becomes possible.

When we point to a question that does appear to belong exclusively to the do- main of stylistics or authorship attribution, is it not always the case that a careful a priori selection of the texts was conducted at an earlier stage of analysis in which authorship and style did impinge on each other? The decision, for example, to un- dertake a nontraditional authorship attribution test necessarily entails never losing sight of the relationship between authorship and style (i.e., genre) since the signal from the latter sometimes "overpowers" the signal of the former (as one sees in Hoover 2013). And when we do find a statistical result that countermands our (literary) expectations, is

not a useful first step to examine the interdependence between authorship and style to account for the surprise?

3. Pastiche Quality and Authorship

Somers and Tweedie's original question can be separated into two: the first relates to the fundamental validity of computational authorship attribution tech- niques and the second relates to a functional definition of what a pastiche is. Ex- plaining these two questions in detail is useful in developing a theoretical perspective to critique and explore Somers and Tweedie's paper as well as for developing a better theoretical foundation for the acceptance or rejection of certain assumptions inherent to the contemporary practice of computational stylistics.

As Somers and Tweedie note, for authorship attribution techniques to be most effective one tracks linguistic habits "which may be the least susceptible to variation" (412)—that is, we look at features that an author does unconsciously since the features an author has no control over are those expected to be least affected by the idiosyncrasies of genre, historical moment, and so forth. Yet if we are seeking to examine the "quality" of a pastiche, as Somers and Tweedie ask, then we are seeking features an author is consciously employing in pursuit of imitating another author. The similarities that are relevant to the literary quality of a pastiche would necessarily be those features for which a human reader is able to readily identify and is likely to be those which a reader has had the most practice at identifying. To be sure, the category of pastiche has perhaps a more overt connection to both past literature written and contemporary culture—it's existence is defined directly by what has already been written and depends upon the reader's recognition of this. These relationships between tradition and culture are perhaps no more or less important to other literary forms, but the category of pastiche specifically asks the reader to reflect upon such relationships directly and overtly.

Reframing Somers and Tweedie's question as two questions allows us to adopt a scheme from R. G. Collingwood's "On the So-Called Idea of Causation" so as to parse the original ambiguity in Somers and Tweedie's into several logically distinct classes. This parsing will allow us to see that attributing the success or failure of authorship attribution algorithms to only the two possibilities of algorithm effec- tiveness or pastiche quality effectively equates two incommensurable ontological systems as if they were logically consistent. To avoid this difficulty, we need only clarify the original question so that we are acting in a logically consistent manner. However this particular scheme comes with a high theoretical cost. To resolve the ambiguity inherent to Somers and Tweedie's question, it may be necessary to re- sort to literary descriptive categories for which the identification by a finite series of computable steps may be theoretically forbidden.

References

- Collingwood, R. G.** (1938). "On the So-Called Idea of Causation." *Proceedings of the Aristotelian Society, New Series* 38: 85-112
- Craig, H. (1999).** "Authorial Attribution and Computational Stylistics: If You Can Tell Authors Apart, Have You Learned Anything About Them?" *Literary and Linguistic Computing* 14: 103-13.
- Erasmus, D.** (1992) *Collected Works of Erasmus*. ed. and trans. by James Brady and John Olin. Toronto: University of Toronto Press.
- Garcia, A. and Martin, J.** (2007). "Function Words in Authorship Attribution Studies." *Literary and Linguistic Computing* 22, No. 1: 49-66.
- Jockers, Matthew.** (2013). *Macroanalysis: Digital Methods and Literary History*. Urbana-Champaign: University of Illinois Press.
- Hoover, D. and Hess, S.** (2009). "An exercise in non-ideal authorship attribution: the mysterious Maria Ward." *Literary and Linguistic Computing* 24(4): 467-89.

- Hoover, D.** (2013). "The Full-Spectrum Text-Analysis Spreadsheet." DH 2013 Conference Abstracts. University of Nebraska-Lincoln; (2003). "Multivariate analysis and the study of style variation." *Literary and Linguistic Computing* 18(4): 34160.
- Love, H.** (2002). *Attributing Authorship: An Introduction*. Cambridge: Cambridge University Press.
- Ramsay, Stephen.** "Toward an Algorithmic Criticism." *Literary and Linguistic Computing* 18.2 (2003): 167-74.
- Somers, H. and Tweedie, F.** (2003). "Authorship Attribution and Pastiche." *Computers and the Humanities* 37: 407-29.

Starting the Conversation: Literary Studies, Algorithmic Opacity, and Computer-Assisted Literary Insight

Plasek, Aaron

New York University, United States of America

Hoover, David L.

New York University, United States of America

1. Probing questions of literary interpretation with computer-aided text analysis

Thirty-seven years after Milic (1966) rejected the untenable fear of many literary scholars that "the study of literature may become mechanical if it is processed by a computer" (4), Ramsay (2003) observes that while computers can successfully pursue "empirical validation of 'impressionistic' or 'serendipitous' critical readings" (173), such uses "have not penetrated the major journals of literary study," and that, "in general, our methods are perceived as some sort of positivistic last stand" (168). Wanting computer-assisted text analysis to have a prominent place in the "wider community of humanist scholars" (2003: 173), Ramsay argues that all literary interpretation is more arbitrary than scholars acknowledge (2003: 170), and that interpretation "must present its alternative text as a legitimate counterpart—even a consequence—of the original" (Ramsay 2011: 56). But, à la McGann (2001), "[a] true critical representation does not accurately (so to speak) mirror its object; it consciously (so to speak) deforms its object" (173). *Reading Machines* echoes this: "The critic who endeavors to put forth a "reading" puts forth not the text, but a new text" (Ramsay 2011: 16). How we put forth "new" texts, McGann argues, deeply depends on our "idiosyncratic relation to the work" (116). Computers help us explore these idiosyncrasies by expanding the number of deformed texts we can produce; reading these "new" texts brings us "to a critical position in which we can imagine things about the text that we didn't and perhaps couldn't have otherwise known" (McGann 2001: 116).

Ellis & Favat (1966) also employ rearrangement-as-interpretative procedure and provide a useful counterpoint to McGann and Ramsay by emphasizing that computers merely "enhance" what literary critics already do (1966: 637). Ellis & Favat use the "opportunity for the different examination and reordering of [Huckleberry Finn]" (637) to ask questions about Huck's speech regarding the relation of "death" and "family" by generating concordances (630-33), but concordances have been used since at least the 13th century. What is new, Ellis & Favat contend, is that the critic can now "ask questions that previously he could have only wished to ask" (638) because the computational efficiency of evidence-gathering allows us to pursue questions previously pragmatically impossible (637). Yet Ellis & Favat suggest that certain kinds of reordering are not valid: "The fact that the computer has aided the scholar does not mean that this critical procedure has been violated" (637). Understanding what would constitute a violation positions us to better appreciate what is new in Ramsay's algorithmic criticism.

When does "the different examination and reordering of data" and the "grouping of [text]" cease to be valid "evidence" (637)?

2. Algorithmic criticism as literary chatbot

Ramsay's textual analysis, grounded in deformance, is a mechanism for destabilizing linguistic and cultural assumptions when reading texts. Although deformance has been a source of contention (Hoover 2005, 2007), many new DH scholars seem to underestimate its importance and implications. We argue that the practice of deformance arises, in part, from a broader conception of the humanities that emphasizes multiplying possible solutions and facilitating interesting and unpredictable discussions rather than finding particular solutions. Thus Ramsay insists that literary criticism should aim not to arrive at the meaning, but to ask "how do we ensure that [a text] keeps on meaning?" (2003: 170). He reaffirms this eight years later: "conclusions are evaluated not in terms of what propositions the data allows, but in terms of the nature and depth of the discussions that result" (2011: 9), echoing McGann's claim that "the critical and interpretative question is not 'what does a poem mean?' but 'how do we release or expose the poem's possible meaning?'" (2001: 108). More recently, surprised by the difference in the words used by male and female speakers in *The Waves*, first discussed in *Reading Machines* (2011), Ramsay asks,

Do we imagine that such further experiments would resolve long-standing questions about gender and language? Do we really want those questions resolved? That last question may seem slightly perverse, but I believe that, in the end, what is most distinct about humanistic inquiry is its resistance toward final answers. It is the goal of the seminar to answer questions, but mostly by proposing them more fruitfully. The humanities wants for itself a world that is more complex than we thought . . . We are in search of a conversation[.] (2012: 11-12)

Ramsay's question is "how successful the algorithms were in provoking thought and allowing insight" (Ramsay 2003: 173), and he is ultimately "more concerned with evaluating the robustness of the discussion that a particular procedure announces" (Ramsay 2011: 17). Unfortunately, emphasizing the "robustness of the discussion" that a deformed text promotes may de-emphasize the algorithms used to generate it. Yet if the algorithm that deforms the original text is to facilitate our interpretive insights, knowing what the algorithm does seems crucially important.

3. Continuing the conversation: getting algorithms out of the black box

Deformance, like computational stylistics, necessarily focuses our attention on certain narratives of meaning, drawing our attention to specific words, phrases, images, and patterns at the expense of others. Lotaria in Calvino's *If on a Winter's Night a Traveler* (1979) "reads" a novel by looking at the words that appear 19 times—namely, "blood, cartridge, belt, commander, do, have, immediately, it, life, seen, sentry, shots, spider," and so forth—and observes that "it's a war novel, all action, brisk writing, with certain underlying violence" (182). She answers the literary equivalent of a factual question using a simple and relatively transparent method. However, when Ramsay lists the most characteristic words used by characters in *The Waves* in order to "participate in [a] literary critical endeavor beyond fact-checking" through the use of the tf-idf algorithm (Ramsay 2011: 10), the resulting "deformed text" he discusses cannot be reproduced by the procedure he describes. This is partly because he does not use the precise tf-idf equation he presents, but (more importantly) because of some interpretative decisions that are by no means obvious. Our point, however, is less to quarrel with Ramsay's decisions than to interrogate his method.

Ramsay (2011) does not explain the exact algorithm that produces his men-only and women-only words, but they are simple enough to identify. Rather than 90 men-only and 14 women-only words, however, we found 117 and 10. The mismatches are caused mostly by Bernard's final retrospective

chapter, which Ramsay has (quite reasonably) omitted from his analysis (though he confirms by email that he should have noted this). Re-analyzing without the final chapter reveals a few remaining discrepancies. For example, Ramsay lists "banker" and "Brisbane" as men-only words, but they appear in Neville's and Bernard's monologues only as imagined quotations from Louis. Discussing this kind of decision, we suggest, could deepen and engage a conversation about *The Waves*.

More significantly, Ramsay's provocative lists of 90 men-only and 14 women-only words rest problematically on the amounts of text by the two genders. Even without the final chapter, there are about 35,000 words by the men and only 20,000 by the women, a discrepancy that "explains" the preponderance of male-only words. To "prove" this one could simply cut each male monologue to the length of its corresponding female monologue. (Corresponding how? Matching longest to longest, shortest to shortest? Why?) We chose a different "deformation," randomizing the monologue lines and cutting each monologue to the length of the shortest, equalizing each character's contribution. (Why?) This deformed text produces 31 women-only and 29 men-only words. Ramsay is right that the algorithm merely begins the argument, but the "provocative" revelation that the men share more words than the women seems deceptively and inappropriately provocative: it rests merely on the lengths of the monologues. (Why are the male monologues longer?) The nature of the male and female words remains provocative and suggestive for a conversation about gender in *The Waves*:

<i>ALL MEN, NO WOMEN</i>	<i>ALL WOMEN, NO MEN</i>
boys	feeling
poet	dropped
letters	waste
weep	swing-doors
office	oppose
wheel	hundred
waistcoat	however
telephone	ease
suffering	board
arms	bernard's
sheer	beak
possible	bag
lord	approach
god	able
friend	
	pavement
	stockings
	shoes
	hide
	step
	front
	settle
	music
	lambert
	fills
	breath
	shot
	real
	pirouetting
	branch
	bedroom
	wash
	wander
	tennis
	soften
	shelter
	million
	matter
	diamonds
	rushes
	pulls
	fling
	cotton
	coarse
	bowl
	antlers

But our deformation's doubling of women-only words raises interesting questions, and warns against over-interpreting the lists. Obviously, men use some of our women-only words in the parts we left out, a problem exacerbated by the rarity of these only-words: the highest frequency above is 4. All this suggests a reconsideration of the initial decision to use tf-idf in the first place. We use this opportunity to argue both that it is imperative that text analysis researchers carefully outline their procedures so that others can reproduce the original results and that more attention be given by the community at large to reexamining earlier results. Framing literary questions algorithmically is beneficial because stating our assumptions in computable terms may reveal our own hidden assumptions about the text we are examining, our assumptions regarding literary interpretation, or both. However, transforming our own literary methods into algorithms is always somewhat imperfect: foregrounding these interpretative decisions rather than hiding them allows the critical conversation to continue in a more valuable fashion by incorporating the difficulties of the algorithm into the act of literary interpretation itself.

References

- Ellis, A. and Favat, F.** (1966). "From Computer to Criticism: An Application of Automatic Content Analysis to the Study of

Literature." In Stone, P., Dunphy, D., Smith, M., and Ogilvie, D. (ed.), *The General Inquirer: A Computer Approach to Content Analysis*. Cambridge: MIT Press, 628-38.

- Hoover, D.** (2007). "The End of the Irrelevant Text: Electronic Texts, Linguistics, and Literary Theory." *DHQ: Digital Humanities Quarterly* 1(2).

_____. (2005). "Hot-Air Textuality: Literature after Jerome McGann." *TEXT Technology* 2: 71-103.

McGann, J. (2001). *Radiant Textuality: Literature after the World Wide Web*. New York: Palgrave.

Milic, L. (1966). "The Next Step." *Computers and the Humanities* 1(1): 3-6. Ramsay, S. (2012). "Textual Behavior in the Human Male." (Revised March 2012 transcript, 14 pages.) *Journal of Digital Humanities* 1(1): 32.

_____. (2011). *Reading Machines: Toward an Algorithmic Criticism*. Urbana: University of Illinois Press.

_____. (2003). "Toward an Algorithmic Criticism." *Literary and Linguistic Computing* 18(2): 167-74.

Who is we? The social media project: Día de las humanidades digitales/Dia das humanidades digitais

Priani, Ernesto

Universidad Nacional Autónoma de México, Mexico

Spence, Paul

Spence, Paul
King's College London

Galina Russell, Isabel

Universidad Nacional Autónoma de México, Mexico

González-Blanco, Elena

Universidad Nacional de Educación Distancia, España

Paixão de Sousa, Maria Clara

Universidade de São Paulo

Alves, Daniel

Universidade Nova de Lisboa

Barrón, José Francisco

Universidad Nacional Autónoma de México, Mexico

Godinez, Marco Antonio

Universidad Nacional Autónoma de México, Mexico

Guzmán, Ana María

Universidad Nacional Autónoma de México, Mexico

The last few years have seen intense debates in the digital humanities community not only in its definition but also its configuration. Several scholars have pointed out that the community is predominantly made up of scholars from a handful of mainly Englishspeaking countries and that linguistic and geographic diversity is sorely lacking. Questions related to ethnicity, gender, race, language and class have been raised within the DH community, as it seeks to find a more global and inclusive organization model. One of the main issues of course, is attempting to integrate with groups of scholars that have not necessarily identified themselves yet as a community or do not even know that DH exists.

Spanish and Portuguese speaking countries, some with long traditions already in *humanidades digitales/humanidades digitais*, have hosted a number of DH events and activities in the last few years, including several conferences and seminars, and recent attempts to build formal networks and associations in these two languages have sought to address concerns regarding international representation and visibility. Therefore the first step in this networkbuilding exercise was to find those who identified themselves as "humanista digital" (Galina and Priani, 2011). As Isabel Galina (2013) pointed out in her keynote presentation at the DH2013 conference: "behind this problem of defining digital humanities (what we are and what

we do) there is an additional now ineludible problem ‘who is we?’”

The DiaHD/DiaHD (*Día de las humanidades digitales*/*Dia das humanidades digitais*) initiative aimed to answer this question. Coordinated by a group of digital humanities organizations and institutions in Spain (Humanidades Digitales Hispánicas), México (RedHD), Portugal (Faculdade de Ciencias Sociais e Humanas, Universidade Nova de Lisboa) and Brazil (Humanidades Digitais, Universidade de São Paulo), with the support of centerNet, the event sought to identify and bring together the work of Spanish and Portuguese speaking digital humanists in Europe and Latin America. The proposal of DiaHD was based on the model of the international centerNetsponsored project *Day in the Life of the Digital Humanities* (*DayofDH*, digitalhumanities.org/centernet/initiatives/) , in which digital humanists all over the world are invited to participate in “a social research project designed to document ‘just what do computing humanists really do?’” (Rockwell et all, 2012). Even though the starting point of the project was effectively almost a direct translation of Day of DH, including the basic question: ‘*¿Qué es lo que hacen realmente los humanistas digitales?*’ / ‘*O que fazem os humanistas digitais?*’, we could not avoid the context in which the project was launched. Behind the question ‘What do digital humanists do?’ we also wanted to know ‘who is this we?’ and ‘where is this we?’. This dual aspect of identification and localization broadens the outlook of DiaHD as a social research project, which then becomes a “process of reflection of what we have created and how it fits in with the socalled global DH community” (Galina, 2014).

II. Execution of the Project

On June 10th 2013 the *Día de las humanidades digitales/Dia das humanidades digitais* took place. The event was managed by two working groups: the international organization group, formed by representatives of the institutions involved, and the local organising group, formed by members of the technological staff of the Facultad de Filosofía y Letras de la UNAM, which managed the technical infrastructure. We were fortunate to count on the generous advice of the creators of the original *Day in the Life of Digital Humanities*, including Geoffroy Rockwell.

Outreach and other activities related to being more inclusive are time consuming. It is important to note that multilingualism requires additional effort both in developing the tool (invitations and instructions for participating) as well as the more qualitative analysis of the data. This DiaHD selected only two languages but we are well aware that others could have been added. If DH is to be successful in expanding the extra effort required should be considered. However, as we believe DiaHD has shown, this relatively small effort pays off considerably in our ability to broaden our definition of ‘we’.

III. Answering the questions

The original *Day in the life of Digital Humanities* was conceived as an individualistic approach to the activities of digital humanists: “The motif [of DayofDH] suggests the documentation of a subject’s ‘real’ life, emphasizing the ordinary aspects of their environment over the extraordinary” (Rockwell et all, 2012). As our project was closely based on DayofDH, we expected that the participants might respond in the same way, as persons documenting the ordinary aspects of their lives. One of the most important results of our experiment was the manner in which the community decided to answer the question *¿Qué es lo que hacen realmente los humanistas digitales?* / *O que fazem os humanistas digitais?* Whereas one part of the community, many with previous experience of participating in the Day in the life of Digital Humanities, followed the model of documenting everyday life, another group chose a different approach to answering the question: they decided to create collective blogs to document the work of their institutions or projects. One relevant example of this is the blog devoted to the humanidades digitais group in Brazil. What does this differing reception of our original invitation signify, and is it meaningful

in defining the “we” in Spanish and Portuguese speaking digital humanities? Does the “we” have a less individualistic nature in some regional and cultural contexts?

For DiaHD 95 blogs were created, out of which 70 were actively used. Of this 57 of the blogs were written in Spanish, whereas 13 were written in Portuguese. As the invitation was based on language and not by region, the geographical location of the authors was not restricted to Spanish and Portuguese speaking countries. We had authors from Argentina, Brazil, Canada, Great Britain, Mexico, Portugal, Spain, Sweden and United States. As far as participating countries was concerned, Spain was the most active with 29 blogs, followed by Mexico with 16, Portugal with 8, and Brazil and USA (both 5). This means that the project integrated researchers from different academic cultures and most likely with different notions of what DH is. We can see this contrast in the blogs created: 37 blogs were collective, and represent projects, magazines, labs, etc. (only 2 of them were located in a non Spanish and Portuguese speaking countries) and 33 were personal blogs.

These results seem to indicate that there are a number of scholars involved in DH projects in what are traditionally underrepresented regions and languages in the DH community. As a social media project, DiaHD points towards the fact that by developing tools in other languages other than English and directly targeting other communities by not relying only on traditional DH communication channels, it is possible to incorporate the experiences of digital humanists that do not usually participate.

V. Conclusions

As a result of a preliminarily observation we might point that the absence of a consolidated, consensual and collective profile for Spanish and Portuguese speaking digital humanities communities is relevant to understanding why many of the participants in DiaHD/DiaHD preferred an institutional voice to a personal one. This may imply, also, a segmentation in the Spanish and Portuguese speaking DH community: some, with previous involvement in the international DH community, identifying themselves as researchers with a specific DH focus, while others prefer to identify DH with work in specific projects, and to maintain their professional identity as a distinct entity. As DH is a field whose definition has largely emerged from a North Atlantic tradition, and responds to a concrete subset of global academic culture, it is clear that there is much work to be done in negotiating the development of the field in other academic traditions, each with their own cultural and practical models. DiaHD/DiaHD not only focused debate on one area of nonAnglophone DH; it also led to a number of practical initiatives continuing the development of Spanish and Portuguese speaking digital humanities, including the mapaHD project (mapahd.org) and the Portuguese speaking association, AHDig (ahdig.org), an initiative that came about as a direct consequence of the participation of Portuguese and Brazilian researchers at DiaHD.

References

- Dacos, Marin.** (2013) “*La stratégie du Sauna finlandais*” in Blogo Numericus, May 2013. blog.homonumericus.net/article11138.html
- Fiormonte, Dominico.** (2012) “*Towards a Cultural Critique of Digital Humanities*”, Historical Social Research – Historische Sozialforschung, Special Issue, no.141, HSR vol.37 2, p.5976. www.cceh.unikoeln.de/files/Fiormonte_final.pdf
- Galina, Isabel y Priani, Ernesto.** *Is There Anybody out There? Discovering New DH Practitioners in other Countries at Digital Humanities Conferences 2011* dh2011.stanford.edu/wpcontent/uploads/2011/05/DH2011_BookOfAbs.pdf
- Galina, I** (2014). *Geographical and linguistic diversity in the Digital Humanities, Literary and Linguistic Computing*, Oxford Journals, 29(3)
- McPherson, Tara** (2912). “*Why are the Digital Humanities so White?*”, in Debates in the Digital Humanities, University of Minnesota Press

Geoffrey Rockwell et all (2012), *The Design of an International Social Media Event: A Day in the Life of the Digital Humanities*. Digital Humanities Quarterly, Volume 6 Number 2, www.digitalhumanities.org/dhq/vol/6/2/000123/000123.html

Dacos, 2013, Fiormonte 2012 and McPherson, 2012

Primer Encuentro de Humanistas Digitales de la RedHD (may 17-18 2012), workshops in Lisbon (November 2011 and June 2013), conference at University of Navarre in May 2013 (www.unav.edu/congreso/humanidadesdigitales), HDH conference (hdh2013.humanidadesdigitales.org), 1st seminar USP, Brazil (October 2013).

RedHD, HDH, AHDig (Associação das Humanidades Digitais)

Digital Humanities conference 2013, dh2013.unl.edu
dh2013.filos.unam.mx

Humanidades Digitais dh2013.filos.unam.mx/
humanidadesdigitaisusp . Other examples of this collective and institutional response are Mexican projects like "Proyecto de investigación sobre métrica y orografía áurea" of the Universidad Autónoma de Ciudad Juárez dh2013.filos.unam.mx/vozyverso/ o the blog, "Libros de Baubo": bitácora de trabajo" dh2013.filos.unam.mx/librosdebaubo . In Spain the blog of the "Red Internacional CHARTA" dh2013.filos.unam.mx/chartha/2013/06/10/redinternacionalcharta/ and the project on "el repertorio métrico digital castellano" dh2013.filos.unam.mx/remetca . In the case of United States we have "The Littera Project"

dhd2013.filos.unam.mx/thelitteraproject/ , and in Portuguese, "máquinas & manuscritos"
dhd2013.filos.unam.mx/maqman .

An initiative led by **Silvia Gutierrez** and **Érika Ortega** which began in Gutierrez's blog
dhd2013.filos.unam.mx/sigutierrez/

Reconstruction and Display of a Nineteenth Century Landscape Model

Priestnall, Gary

gary.priestnall@nottingham.ac.uk

1School of Geography, the University of Nottingham, United Kingdom

Katharina, Lorenz

Digital Humanities Centre, Department of Classics, the University of Nottingham, United Kingdom

Mike, Heffernan

1School of Geography, the University of Nottingham, United Kingdom

Joe, Bailey

1School of Geography, the University of Nottingham, United Kingdom

Craig, Goodere

Digital Humanities Centre, Department of Classics, the University of Nottingham, United Kingdom

Robyn, Sullivan

Digital Humanities Centre, Department of Classics, the University of Nottingham, United Kingdom

Introduction

Physical landscape models have been used for hundreds of years as a means of offering people privileged overviews of landscapes, allowing them to effortlessly appreciate spatial relationships between places of interest. The role of digital technology for capturing, preserving and analysing existing models has been demonstrated by Niederoest (2002) using photogrammetric techniques, focussing on the model constructed by Franz Ludwig Pfyffer in the late eighteenth century. This paper will describe the use of digital scanning, processing and 3D printing for exploring a landscape model that no longer exists but where a large number of negative

moulds remain. It offers an example of the use of capture and presentation technologies to produce an exhibit from material that would otherwise be inaccessible to audiences. The specific case study is a physical landscape model of the English Lake District created by Thomas and Henry Mayson in 1875 which gave visitors to the town of Keswick, Cumbria an unprecedented view of the landscape they were about to explore. Huge efforts went into creating this model which claimed to faithfully represent the contours and other details of the Ordnance Survey maps which had recently been surveyed but which were not commonly used by the public at that time for recreational purposes. The model is believed to have been displayed until the 1960s but all that now remains are some mouldings created 'for future use' (found in storage in 2012) along with some other material including posters, the commissioning letter and some original map sheets (Figure 1).



Fig. 1: Remaining archival material relating to Mayson's Ordnance Model of 1875.

The Mayson model was larger and more detailed than others displayed at that time and became a popular tourist attraction. A poster advertising the model reads "*The model has been constructed mathematically from the Ordnance Survey... Parties visiting this model will see the correct topography of the Lake District, and can thereby readily plan either long or short excursions as time will permit. They will also gain a better idea of the whole of the Lake Country than is to be obtained from any other source*".

Background

Whilst very effective cartographic relief representations were available in the nineteenth century, particularly in alpine regions as described by Collier et al (2003), the majority of visitors to the English Lake District may at best have used a guidebook containing small maps but were unlikely to have had a good appreciation of the landscape setting of the town they were visiting. Physical landscape models have often been used in public settings to help convey a sense of spatial context for visitors, using geographical features such as mountains, lakes, buildings and roads to provide a frame of reference, and Keswick in the nineteenth century was no exception. Whilst some models from this period still exist, including the Flintoft model made in 1834, several of the larger models including that of the Mayson brothers, have been destroyed. In the case of the Mayson model there is an opportunity to reconstruct parts of the model due to the discovery of a set of negative moulds. The use of laser-scanning technology as reported by Terdiman (2012) in relation to the Smithsonian archive clearly offers the fidelity of capture necessary to extract the detail present in the Mayson moulds. In addition to scanning the moulds, digital geo-processing and 3D prototyping will enable us to explore the process of model building and the relationships between the model and the maps on which it is said to have been based. This broader view on the potential for physical fabrication to add richness to digital humanities investigations is seen in Elliot et al (2012) and Sayers et al (2013).

Research challenges

Through the development of a workflow combining scanning, processing and physical fabrication the project explores the technical aspects of physical model building and in particular the relationship between cartographic survey and physical model construction. The historical context of the model as part of the visitor experience, before Ordnance Survey mapping

saw popular recreational use, is of particular interest. A great challenge is to apply technology in a creative manner in order to re-present the collection of objects in a form that is appropriate for public viewing, conveying both the fidelity and scale of the original model, but also advancing our understanding of the process of model building and the original context of its use as an informative visitor attraction.

Digital reconstruction

After recovering and cleaning the collection of moulds each item was scanned using a FARO Laser ScanArm V3 and the resulting point clouds tidied within the PolyWorks software. A program was written to convert each point cloud into a form suitable for processing within the ArcGIS Geographical Information System (GIS) package, including the inversion of the vertical dimension. The points were then processed to produce continuous Digital Surface Models from which hillshaded images were derived to assist in locating each mould for the purpose of geo-referencing, as the location and orientation of most of the moulds was not known. From initial processing within the GIS there is some evidence of slight vertical exaggeration when comparing the model against modern elevation data however this will need a full and systematic study and will form part of the ongoing research agenda. Selected tiles were sent to a CMS Athena 5-axis CNC milling machine to produce positive reconstructions of the relevant mould. The generalised workflow is shown in Figure 2.

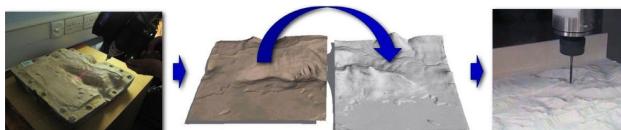


Fig. 2: Digital Reconstruction workflow.

Designing a new visitor experience

The knowledge gained about the Mayson model both in terms of the process of construction and the context of its display will form part of a public display in Keswick museum from May 2014. The challenge for re-presenting this collection is to convey the detail of the original model, the process through which the original model was made and how this relates to modern forms of survey, mapping and model building, helped by the active involvement of the Ordnance Survey, on whose maps the original model was based. Figure 3 shows some key elements of the display, including an example of an original mould, a 3D fabrication of the scanned data from that original mould with map data projected down onto it (Priestnall et al. 2012) including modern cartographic representations but also the 1860s map (Figure 3, centre), attempting to emphasise the data from which a model has been derived but also how it could be re-presented and re-interpreted (Lorenz, Schofield and Noond, 2006). The scale and context of display of the original model will also be presented (below right) along with a temporary installation featuring ten fabricated tiles distributed around the gallery space in their true to-scale geographical positions to convey the size of the original model.

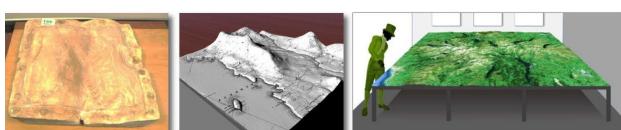


Fig. 3: Re-presenting the collection

Conclusion

Digital technologies have allowed us to explore a nineteenth century physical landscape model that no longer exists. We have gained an insight into its creation, its detail and accuracy, and the importance of its role in the visitor experience at a

time when maps were less commonly used for recreational purposes. We have developed workflows to combine 3D scanning, digital processing, 3D fabrication and projection to produce an exhibit from historical objects that would otherwise be inaccessible to the general public. In this case not only has the technology been of value to preserve a collection but it has enabled us to digitally reconstruct new forms of information from the artefacts that remain. Particular issues arising from the need to apply geographic coordinate information to the scanned objects were handled within a GIS and work is ongoing to compare the historic terrain model with modern digital equivalents. The overall findings resonate with ongoing research into the development of projection-enhanced physical landscape models for public display and their power for promoting an awareness of spatial context in viewers of those models. The techniques could also contribute to more general design guidelines for using technology to promote an awareness of spatial context and to support the process of interpretation and reconstruction where the geographic landscape model is the backdrop but not necessarily the focus of attention.

Acknowledgements

Charlotte Stead and Pat Maskell at Keswick Museum; Nikki Tofts at Helena Thompson Museum; Glen Hart, Head of Research at the Ordnance Survey; Sarah Beardsley, Scott Wheaver and James Hazzledine at the Centre for 3D Design, School of Architecture and the Built Environment, the University of Nottingham; Sally Bowden at the Centre for Advanced Studies, the University of Nottingham; Frank Priestnall; and Ian Conway, School of Geography, the University of Nottingham. This work has been supported by the AHRC Creative Economy Knowledge Exchange Project "Archives, Assets and Audiences".

References

- Collier, P., Forrest, D., and Pearson, A. (2003)** *The Representation of Topographic Information on Maps: The Depiction of Relief* The Cartographic Journal Vol. 40 No. 1, 17-26
- Elliot, D., MacDougall, R., and Turkel, W.J. (2012)** *New Old Things: Fabrication, Physical Computing, and Experiment in Historical Practice*, Canadian Journal of Communication Vol 37 (2012) 121-128
- Lorenz, K., Schofield, D. and Noond, J. (2006)** *Showing Seeing : In Search of a Graphic Language for the Digital Reconstruction of Ancient Greek and Roman Sites and Monuments*, Digital Resources for the Arts and Humanities Conference, Dartington, UK, 9-12 September 2006.
- Niederoest, J. (2002)** *Landscape as a historical object: 3D Reconstruction and Evaluation of a Relief Model from the 18th Century* The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Vol. XXXIV, Part 5/W3
- Priestnall, G., Gardiner, J., Durrant, J. and Goulding, J., (2012)** *Projection Augmented Relief Models (PARM): Tangible Displays for Geographic Information* In: Proceedings of Electronic Visualisation and the Arts (EVA 2012). 180-187
- Sayers, J., Boggs, J., Elliott, D., and Turkel, W.J. (2013)** *Made to Make: Expanding Digital Humanities through Desktop Fabrication*, Proceedings of Digital Humanities 2013, University of Nebraska.
- Terdiman, D. (2012)** *Smithsonian turns to 3D to bring collection to the world*. CNet. [Accessed November 1st 2013] news.cnet.com/8301-13772_3-57384166-52smithsonian-turns-to-3d-to-bring-collection-to-the-world

Constructing Scientific Archives that Support Humanistic Research

Prom, Christopher

chris.prom@gmail.com

University of Illinois at Urbana-Champaign

1. Introduction

The results of what has been termed the scientific method infuse every aspect of modern life. Science, as a contested enterprise, affects and is affected by political, economic, technological, social, cultural, religious, and ethical factors. Sensational cases illustrate that people are interested scientific data and the process that produces scientific knowledge.¹ Scientific data, and evidence about how it was created, shaped, and interpreted, serve as resources for humanistic studies in disciplines including journalism,² anthropology,³ history,⁴ medicine,⁵ and literary studies.⁶

If scientific information is to be useful for current and future scholarship, we must answer three fundamental questions: (1) Which information is preserved and made accessible? (2) What evidence does it provide? and (3) How can it be used? Future users of scientific records will provide conflicting answers: Archivists cannot anticipate the precise nature of future research, and selecting or arranging materials with the expectation of one particular use may preclude transformative uses.

If the goal of preservation is to make accessible a record of science that is amenable to disparate analyses and interpretations, information professionals must identify and preserve evidence regarding the process of scientific production, not just the results of the process. Given that each lab operates differently, preservation programs must use interdisciplinary modes of analysis and techniques, as must users of archives.

The fields of digital humanities, digital curation, data curation, and digital preservation have long been interdisciplinary, and are becoming increasingly hybrid in terms of the disparate theoretical frameworks and tools of analysis that they utilize. It may well be true that the lines between archives, libraries, digital humanities, and publishing are becoming more fluid, yet we believe that archives are well positioned to serve as centralized sites where these intersections lead to greater collaboration and knowledge dissemination.⁷

2. Academic Archives (AA)

Academic and research institutions have traditionally played a leading role in preserving the 'papers' (i.e. personal archives) of faculty, including scientists.⁸. Archival guides devoted to scientific and technical documentation emphasize that reports, correspondence, photographs, scrapbooks, lab records, and supplementary documentation hold as much continuing research value as formal research products (publications and processed datasets). Ideally, archivists act upon the basis of the core archival doctrine, *respect des fonds*: the idea that grouping records by the function or activity that led to their creation best preserves their value as evidence.⁹ To accomplish this, archivists draw upon concepts and techniques from other disciplines, to protect the records' provenance and original order.

3. Digital Preservation (DP)

DP's core concepts extend archival doctrine and practice. DP establishes requirements and processes for maintaining authentic and trustworthy digital objects.¹⁰ The Open Archival Information System Reference Model (OAIS) and Trusted Digital Repositories Framework provide recommendations concerning system design, policy development, and institutional commitments.¹¹ Research regarding digital personal papers (including faculty archives) recommends particular acquisition, arrangement, descriptive, and access practices that preserve contextual information regarding the creation and use of digital records, such as documentation about the research

environment and the use of communication/dissemination technologies.¹²

4. Data Curation (DC)

DC techniques complement DP and help scholars manage datasets of continuing informational value, by making them preservable, reusable, and computationally reproducible;¹³ and by developing "high-functioning" metadata.¹⁴ While DC provides a rich data-preservation toolkit, it focuses more attention on data than on ancillary documentation, which provides evidence of the ways in which that data was created, used, or interpreted. If the goal is to document the research environment and process, DC methods are but one (albeit, crucial) element in a broader archival strategy.

5. Anthropology

Documenting scientific activity, one of many social arenas in which knowledge is constructed, is best accomplished by direct observation. In the absence of ethnography, archives play an important documentary role by providing insight into the context underlying scientific fact creation. Anthropology's focus on praxis and its reflexive methodological and epistemological underpinnings offer meaningful points of departure for archivists seeking to capture a nuanced record of scientific processes and activities.¹⁵

One goal of science is to produce authoritative, published documents: a materialized fact.¹⁶ The archival challenge lies in documenting social factors at play from data generation through the various stages of processing and interpretation, to the final "point of stabilisation."¹⁷ Archives can lay bare the social factors that are all too easily stripped out when the fact is reified, allowing us to read against and along the scientific grain.¹⁸ As anthropological surrogates, archives must capture the processes and events that destroy and create facts.¹⁹

6. Digital Humanities (DH)

Network analysis, text mining, and information visualization provide an algorithmic supplement to the pursuit of past, present, and future research questions, enhancing analysis of digital objects.²⁰ Recent work with digitized historical records demonstrates how the algorithmic approach offers a lens with which to gaze upon and ascertain meaning from large and unwieldy bodies of data.²¹ As the corpus of materials amenable to computational analysis grows in archives throughout the world, approaches of this type are becoming indispensable.

While great promise lies in the analytic potential of the DH, its practitioners are increasingly aware of the challenge to making the results of their research preservable and in the optimal case, reusable.²² For their part, the familiarity that archivists have with the lifecycle of data prepares them well for having conversations with digital humanists.²³ Archivists can become digital humanists themselves as they utilize approaches like topic modeling to enhance the ways in which they, and their users, interact with archival materials.

7. General Application

An archival processing strategy that is informed by anthropological principles and that uses DH tools will facilitate reflexivity between the archivist's professional practices and the scholar's interpretive possibilities. Evidence of the knowledge production process can be better preserved by integrating DP, DC, and DH concepts and tools as essential elements of a systematic archival processing workflow covering analog, born-digital, and digitized records.²⁴ Archivists should be particularly attuned to six elements of scientific work: inscription, circumstantiality, noise, conflict, credibility, and reification.²⁵

To document these factors, archivists must modify how they appraise, preserve, arrange, and describe scientific records. The challenges to the archival task are many, and include:

- Recognizing the potential of archival sources as anthropological surrogates.
- Weighing the value of records in terms of the insight they reveal in terms of the six elements of scientific work that are listed above.
- Using tools that preserve authenticity, while also providing scholars the ability to understand how the scientific process played out within the social networks and environment supporting scientific research.
- Controlling costs and sustaining the archives over time.

8. Case Study

The strategy being used by the University of Illinois at Urbana-Champaign is illustrated by our ongoing work with the records of Carl Woese (1928-2012), 2003 winner of the Crafoord Prize in Biosciences.²⁶ While a full description of the project is beyond the scope of an abstract, we are integrating DP, DC, and DH concepts and tools into three particular points in our workflow.²⁷

Appraisal/Acquisition

- Professionally photograph Woese's laboratory to document microsocial environment.
- Transfer materials while maintaining original order and recording original placement in lab.
- Create forensic image of Woese's laptop to capture records in a consistent, verifiable manner and avoid unintended data modification.²⁸
- Extract user files with disk analysis reports to use as surrogate during processing and topic modeling.
- Used topic modeling and network analysis tools to help develop processing plans.

Processing

- Arrange analog records into functional series representing Woese's activities.
- Preserve reprint files in original order, including correspondence regarding Woese's methods and conclusions.
- Generate preservation metadata for born-digital content, including genomic datasets.²⁹
- Use open-software and to identify and remove private/confidential records, and to identify documents speaking to the six factors, and to assist in generation of access copies, including topic model as alternate access point.³⁰

Access

- Create a summary online description for all analog and digital files, with file-level inventory.³¹
- Use file conversion and normalization tools to create access copies of the digital and digitized files.
- Provide access copies in zip format, facilitating application of data analysis tools by scholars.
- Present topic-modelled view of data as alternative access point.³²
- Deposit preserved records in Library's digital preservation repository (Medusa), ensuring long term integrity and accessibility.³³
- Undertake migration assessment and planning for at risk file formats.

The process described above is surprisingly cost effective when integrated into the "More Product, Less Process" framework for achieving archival efficiency (more information and examples will be provided at the conference and in a full paper, if selected for inclusion in the proceedings).³⁴

9. Conclusion

In the Woese project, we seek to test one means of preserving evidence about knowledge production in scientific

archives. We believe that techniques like those described above can and should be applied to the archives of important scientists, if we wish for those archives to support humanistic research, but--as with all areas of human knowledge--we realize that our own conclusions are subject to revision after being viewed in the bright light of experience.

References

1. Anthony A. Leiserowitz, Edward W. Maibach, Connie Roser-Renouf, Nicholas Smith, and Erica Dawson (2013), *Climategate, Public Opinion, and the Loss of Trust*, American Behavioral Scientist 57:6: 818-837.
2. Jeff Gerth and T. Christian Miller (2013), *Use Only as Directed*, Pro Publica, accessed October 31, 2013, www.propublica.org/article/tylenol-mcneil-fda-use-only-as-directed.
3. David Zeitlin (2012), *Anthropology in and of the Archives: Possible Futures and Contingent Pasts. Archives as Anthropological Surrogates*, Annual Review of Anthropology 41: 461-80; for a popular account using scientific archives, see David Grann, *The Lost City of Z: A Deadly Tale of Obsession in the Amazon* (New York: Vintage Books, 2010).
4. Harry Woolf, *Manuscripts and the History of Science*, ISIS 53 (March 1962): 3; Lillian Hoddeson, True Genius: The Life and Science of John Bardeen, the Only Winner of Two Nobel Prizes in Physics, (Washington, D.C.: Joseph Henry Press, 2002).
5. Ezra Susser, Hans W. Hoek, Alan Brown, *Neurodevelopmental Disorders After Prenatal Famine: The Story of the Dutch Famine Study*, American Journal of Epidemiology 147:3 (1998): 213-216. We are indebted to Tom Nesmith for this reference.
6. Lisa Yazeck, *Narrative, Archive, Database: The Digital Humanities and Science Fiction Scholarship*, The Eaton Journal of Archival Research in Science Fiction 1:1 (April 2013): 8-13; Susan Haack, "Science, Literature, and The Literature of Science," The Humanities and the Sciences, American Council on Learned Societies Occasional Papers No. 47, 1999, accessed October 31, 2013, archives.acls.org/op/op47-4.htm.
7. Tanya Clement, Wendy Hagenmaier, and Jenny Levine Kries, *Toward a Notion of the Archive of the Future: Impressions of Practice by Librarians, Archivists, and Digital Humanities Scholars*, The Library Quarterly 83 (April 2013): 112-130.
8. Maynard J. Brichford, *University Archives: Relationships with Faculty*, The American Archivist 34, no. 2 (April 1, 1971): 173-181; William J. Maher, *The Management of College and University Archives* (Metuchen, N.J.: Society of American Archivists and Scarecrow Press, 1992), 27-28; Tom Hyry, Diane Kaplan, and Christine Weideman, "Though This Be Madness, Yet There Is Method in 'T': Assessing the Value of Faculty Papers and Defining a Collecting Policy," *The American Archivist* 65, no. 1 (April 1, 2002): 56-69; Tara Zachary Lavar, "In a Class by Themselves: Faculty Papers at Research University Archives and Manuscript Repositories," *The American Archivist* 66, no. 1 (April 1, 2003): 159-196.
9. James O'Toole and Richard J. Cox, *Understanding Archives & Manuscripts. Archival Fundamentals Series*. (Chicago, IL: Society of American Archivists, 2006), 87-131.
10. Joan K. Haas, Helen W. Samuels, and Barbara Trippel Simmons, *Appraising the Records of Modern Science and Technology: A Guide* (Massachusetts Institute of Technology, 1985); Maynard J. Brichford, *Scientific and Technological Documentation; Archival Evaluation and Processing of University Records Relating to Science and Technology* (Urbana, IL: University of Illinois, 1969), 13-15.
11. InterPARES Project, *The Long-term Preservation of Authentic Electronic Records: Findings of the InterPARES Project*, no date. www.interpares.org/book/index.cfm ; Luciana Duranti, *Preservation of the Integrity of Electronic Records. The Archivist's Library v. 2.* (Dordrecht; Boston: Kluwer Academic, 2002).
12. Consultative Committee for Space Data Systems, *Reference Model for an Open Archival Information System (OAIS)*, January 2002, accessed October 31, 2013,

- public.ccsds.org/publications/archive/650x0b1.pdf ; RLG/OCLC Working Group on Digital Archive Attributes. Trusted Digital Repositories: Attributes and Responsibilities (Mountain View, CA: Research Libraries Group, 2002), accessed October 31, 2013, www.oclc.org/research/activities/past/rig/trustedrep/default.html.
12. **Paradigm Project.** *Workbook on Digital Private Papers*, 2005, accessed October 31, 2013, www.paradigm.ac.uk/workbook/index.html ; Jeremy Leighton John with Ian Rowland, Peter Williams, and Katrina Dean, *Digital Lives: Personal Digital Archives for the 21st Century: An Initial Synthesis* (The British Library, 2009), accessed October 31, 2013, britishlibrary.typepad.co.uk/files/digital-lives-synthesis01a.pdf ; AIMS Work Group. AIMS Born-Digital Collections: An Inter-Institutional Model for Stewardship, 2012, accessed October 31, 2013, www2.lib.virginia.edu/aims/whitepaper/AIMS_final.pdf. Tracey P. Lauriault, Barbara L. Craig, D. R. Fraser Taylor, and Peter Pulsifier, "Today's Data Are Part of Tomorrow's Research: Archival Issues in the Sciences," *Archivaria* 64 (Fall 2007): 123-179.
13. **Esther Conway, David Giaretta, Simon Lambert, and Brian Matthews.** *Curating Scientific Research Data for the Long Term: A Preservation Analysis Method in Context*, *International Journal of Digital Curation* 6, no. 2 (July 26, 2011): 38-52, accessed October 31, 2013, doi:10.2218/ijdc.v6i2.204; Anne E. Thessen and David J. Patterson, "Data Issues in the Life Sciences," *ZooKeys* no. 150 (November 28, 2011): 15–51, accessed October 31, 2013, doi:10.3897/zookeys.150.1766; Victoria C. Stodden, "Reproducible Research: A Digital Curation Agenda" (2011), accessed October 31, 2013, academiccommons.columbia.edu/download/fedora_content/download/ac:147764/CONTENT/IDCC-Dec62011-STODDEN.pdf.
14. **Carole L. Palmer, Nicholas M. Weber, Trevor Munoz, and Allen H. Renear.** *Foundations of Data Curation: The Pedagogy and Practice of 'Purposeful Work' with Research Data*, *Archive Journal* no. 3 (Summer 2013), accessed October 31, 2013, www.archivejournal.net/issue/3/archives-remixed-foundations-of-data-curation-the-pedagogy-and-practice-of-purposeful-work-with-research-data/.
15. **Elisabeth Kaplan.** 'Many Paths to Partial Truths': *Archives, Anthropology, and the Power of Representation*, *Archival Science* 2 (2002): 209-220.
16. **Bruno Latour and Steve Woolgar.** *Laboratory Life: The Construction of Scientific Facts* (Princeton: Princeton University Press, 1986), 50.
17. Ibid, 175-176.
18. See **Nupur Chaudhuri, Sherry J. Katz, and Mary Elizabeth Perry, eds.**, *Contesting Archives: Finding Women in the Sources* (Urbana: University of Illinois Press, 2010) **Ann Laura Stoler**, *Along the Archival Grain: Epistemic Anxieties and Colonial Common Sense* (Princeton: Princeton University Press, 2009) and **Caroline B. Brettell**, *Archives and Informants: Reflections on Juxtaposing the Methods of Anthropology and History*, *Historical Methods* 25, no. 1 (Winter 1992): 28-36.
19. **David Zeitlin.** *Anthropology in and of the Archives: Possible Futures and Contingent Pasts. Archives as Anthropological Surrogates*, *Annual Review of Anthropology* 41 (2012): 461-80.
20. **D. Sculley, and Bradley Pasenek.** *Meaning and mining: the impact of implicit assumptions in data mining for the humanities*, *Literary and Linguistic Computing*. no. 4 (2008): 409-424, accessed October 31, 2013, doi:10.1093/linc/fqn019
- David Blei**, *Probabilistic Topic Models*, *Communications of the ACM* no. 4 (2012): 77-84, accessed October 31, 2013, doi:10.1145/2133806.2133826
- Andrew Goldstone and Ted Underwood**, *What Can Topic Models of PMLA Teach Us About the History of Literary Scholarship?* *Journal of Digital Humanities*. no. 1 (2012), accessed November 1, 2013, journalofdigitalhumanities.org/2-1/what-can-topic-models-of-pmla-teach-us-by-ted-underwood-and-andrew-goldstone/.
21. **Robert Nelson**, *Mining the Dispatch*. University of Richmond, dsl.richmond.edu/dispatch/. Elijah Meeks. Karl Grossner, "Developing Kindred Britain". Stanford University kindred.stanford.edu/notes.html?section=originating.lab.softwarestudies.com/p/imageplot.html
22. **Julia Flanders, Trevor Munoz**, *An Introduction to Humanities Data Curation*. guide.dhcuration.org/intro
23. **Alex H. Poole**, *Now is the Future Now? The Urgency of Digital Curation in the Digital Humanities*, *Digital Humanities Quarterly* 7 (2013), accessed March 5, 2014, digitalhumanities.org/dhq/vol/7/2/000163/000163.html.
24. **J. Gordon Daines III**, *Processing Digital Records Manuscripts*, in *Archival Arrangement and Description*. eds. Christopher J. Prom and Thomas J. Frusciano. *Trends in Archives Practice Series* (Chicago: Society of American Archivists, 2013), 87-144.
25. **Latour and Woolgar** (1986), 236-244.
26. **Carl Woese**, *Wikipedia*, accessed October 29, 2013, en.wikipedia.org/w/index.php?title=Carl_Woese&oldid=579341488.
27. See **Christopher Prom**, *Making Digital Curation a Systematic Institutional Function*, *International Journal of Digital Curation* 6, no. 1 (August 3, 2011): 139-152, accessed October 31, 2013, doi:10.2218/ijdc.v6i1.178; links from *Staff Resources*, University of Illinois Archives website, archives.library.illinois.edu/staff-resources/.
28. **Forensic Toolkit Imager** (FTK Imager): www.forensicswiki.org/wiki/FTK_Imager Bitcurator: www.bitcurator.net
29. **NARA File Analyzer and Metadata Harvester**: https://github.com/usnationalarchives/File-Analyzer. DROID: www.nationalarchives.gov.uk/information-management/our-services/dc-file-profiling-tool.htm ; TreeSize Pro: www.jam-software.com/treesize ; Karen's Directory Printer: www.karenware.com/power-tools/ptdirprn.asp.
30. **Thomas Padilla**, *Topic Modeling Archival Materials*, Practical E-Records Blog, accessed November 1, 2013, e-records.chrisprom.com/topic-modeling-archival-materials ; Archon: archon.org
- Our current descriptive catalog/software is *Archon*: archon.org ; we are migrating to *ArchivesSpace*, archivesspace.org ; A description of the Woese Papers is available at archives.library.illinois.edu/archon/index.php?p=collections/controlcard&id=11138
32. **David Blei, Andrew Ng, and Michael Jordan**, *Latent Dirichlet Allocation*, *Journal of Machine Learning Research* (2003): 993-1022; **Robert Nelson**, University of Richmond, *Mining the Dispatch*, accessed October 31, 2013, dsl.richmond.edu/dispatch ; **Lisa Rhody**, *Topic Modeling and Figurative Language*, *Journal of Digital Humanities*. no. 1 (2012), accessed October 31, 2013, journalofdigitalhumanities.org/2-1/topic-modeling-and-figurative-language-by-lisa-m-rhody
33. medusa.library.illinois.edu .
34. **Mark Greene and Dennis Meissner**, *More Product, Less Process: Revamping Traditional Archival Processing*, *The American Archivist* 68, no. 2 (Fall/Winter 2005): 208-263.

Digital Linguistic Archive of the Dutch East India Company (VOC): Modeling a community-sourcing platform for historical linguistic research

Pytlowany, Anna

NUI Maynooth, Ireland ; a.pytlowany@gmail.com
University of Amsterdam, The Netherlands

1. Overview

The primary aim of the project "Digital Linguistic Archive of the Dutch East India Company (VOC)" is to bring together in one online platform all Dutch documents from 1600-1825, which were written either by VOC employees or under the auspices

of the Company, and relate to the languages of the "newly discovered" territories, mainly South East Asia and Africa.

The secondary, yet more ambitious goal is to create an online platform enabling researchers to contribute and exchange knowledge about the source documents, possibly including social editions of some texts. Such collaboration would bring together researchers from different countries and languages, and result in opening access to a vast range of unpublished linguistic material.

2. Source material

The archives of the Dutch East India Company (VOC) are spread between The Hague, Jakarta, Kaapstad, Colombo, Madras and London. Their historical and cultural significance was recognized by UNESCO in 2003, when they were added to UNESCO's Memory of the World Register, a 'World Heritage List' for the preservation of valuable archives and library collections. Considering the size of the VOC archives – comprising 34 million pages in total, including 1.4 kilometers of shelf material in the Hague and 2.5 km in Jakarta – one can only be disappointed at the scarcity of available documents relating to the languages of the Dutch maritime empire.

However, a closer inspection of various independent library records yields surprising finds: some of these Dutch manuscripts and printed books may be found scattered in private and public collections from Paris, to London, to Venice, to Sydney. Up to this point, only a few isolated studies on these manuscripts have been available. An online database detailing all bibliographical information available would help unravel the documents' provenance, itineraries, and interconnections.

So far, these documents have never been assembled and arranged into one comprehensive collection. Compiling such an inventory would be highly desirable because it would allow a better evaluation of the scope and content of Dutch colonial linguistic heritage. Last but not least, online publishing of digital editions would open access to a vast range of previously unpublished linguistic material.

3. The need for collaboration

One of the most significant challenges to face in the modeling of this project is the potential multitude of languages and scripts involved. Although the language of the main body of texts is predominantly Dutch, the grammars and vocabularies, by their very nature, each contain at least one other "exotic" language, often written in the native, non-Latin script. Inasmuch as one part of the multi-layered text may be accessible to a particular researcher, chances are that the remaining levels require further work to make them readily understandable.

This is perhaps the most compelling characteristic of the archival material of this type: it lends itself perfectly to a collaborative, crowdsourced (or rather: community-sourced) online undertaking. The tasks may include: adding new documents, transcribing or translating existing documents, proofreading and validating transcriptions, making annotations, and adding references and annotations.

Let's demonstrate it on a real-life example: a newly discovered and digitised 17th century treatise on Tamil letters, recently made available online by the Utrecht University Library (Ms. 1479):

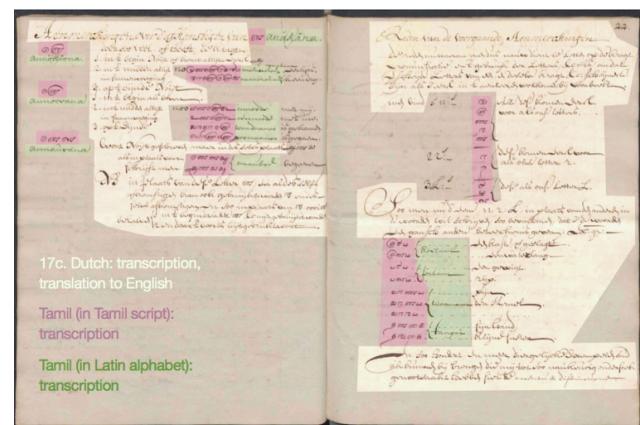


Fig. 1: A newly digitized manuscript: user story

A famous researcher on Tamil would be quite interested in studying it, but she does not know Dutch, or maybe is not familiar with the 17th century Dutch paleography. However, a paleography student may find it motivating and worthwhile to practice his skills on a real-life manuscript; a retired English teacher from Holland may then contribute the English translation.

Once the text is rendered into English, it opens new possibilities for a wide community of non-Western scholars who may be interested in early descriptions of their native languages. They, in turn, can contribute their transcription of parts written in non-Latin scripts, as well as annotations regarding the linguistic content of the documents.

3. Content organisation

The sources will be indexed by title, author, language, printer / publisher (where applicable), and the relevant linguistic category. This will enable a dynamic visualization of interrelations between any selected two criteria by means of relationship graphs, to help users understand the networks and collaboration patterns.

Additional tools, such as geo-referencing, annotations, etc. can also be developed. For the purposes of paleographic research and comparison, a sample of the handwriting from the text would also be provided.

4. Challenges

However, before any IT solutions can be developed, other key issues will have to be addressed, notably in relation to copyrights and intellectual property rights. How can the existing digitized object from different libraries be brought together? How to ensure access to documents behind a paywall? Could the added value of knowledge and online traffic act as trade-off for libraries?

The other methodological questions will concern ways of organising and managing the crowdsourcing community based on previous experience from comparable projects. The issues of authority, access and quality control will have to be addressed in order to ensure adherence to rigorous scholarship standards.

On automatically disambiguating end-of-line hyphenated words in French texts

Pöckelmann, Marcus

marcus.poeckelmann@informatik.uni-halle.de

Institute of Computer Science, Martin-Luther-University Halle-Wittenberg,
Germany

Ritter, Julia

julia.ritter@student.uni-halle.de

Institute of Romance Studies, Martin-Luther-University Halle-Wittenberg,
Germany

Gießler, André

andre.giessler@informatik.uni-halle.de

Institute of Computer Science, Martin-Luther-University Halle-Wittenberg,
Germany

In printed texts usually a lot of words are separated by a hyphen at line breaks. Such a hyphenation is made if the last word is too long for the current line particularly with regards to a justified text alignment. Whereas in many cases an additional hyphen (soft hyphen) will be appended to the first part of a word, some words already contain a hyphen (hard hyphen) that can be used for the line break. During different steps of automated text processing hyphenation can be hindering if the correct spelling of a word, whether with or without hyphen, is unknown. Just think of applications in which the text shall be annotated automatically or shall receive a different typesetting. In such cases it is desirable to use a self-acting or at least a semiautomatic approach in contrast to make manual decisions for every word's correct spelling, which can be notably time-consuming for long texts.

There is only a sparse amount of comments in the literature how to handle the problem described above, especially in French. Some publications propose to make the decisions manually^{1,2}. The documentation of the Oxford Concordance Program³, which is a software from the 1980s, states that it "has a facility to request that hyphenated words at the ends of lines should be reconstituted" but without giving details of the realization of this feature. One trivial procedure, removing all end-of-line hyphens, is used in a paper about tokenization⁴. This paper also mentions the use of a dictionary as possibility to reduce the error rate, which is an essential part of our approach discussed later.

Simply joining the separated parts of a word by leaving the hyphen out may solve the problem for most instances in many languages, e.g. English or German. In French however a more complex approach is necessary because the hyphen is frequently used in positions other than the end of a line. Particularly this includes the building of compounds with prefixes, nouns and pronouns as well as numbers that made their way into the written language. Whereas hyphenation in French usually follows well-defined rules nowadays, these rules changed through time and had not been applied consistently. Thus a reliable rule-based approach for disambiguating end-of-line hyphenated words is unlikely.

To solve the challenge we have developed a dictionary-based technique for reversing the hyphenation for a given French text. The approach consists of three steps. First, an internal attribution is computed which determines the number of occurrences for every word of the text under consideration. Thereby only occurrences not separated by a line break are considered. Thus the text itself will become a reference for the correct spelling of a given separated word by comparing the number of occurrences of both possible spellings. The second step is a query in an external dictionary. Again both spellings of the word, whether with or without hyphen, are searched in the given dictionary. The third step merges the information of the two previous steps in order to provide a guess for the correct spelling. Both of the previous steps may have led to either no indication at all, or to an indication for exactly one spelling or to an indication rendering both spellings probable. Thereby 16 cases are possible. Our approach assumes that the spelling with hyphen is correct if the internal attribution returns only the entry for the spelling with hyphen even if the external dictionary says differently. This keeps the consistent spelling of the author or the age of the text. The hyphen is also chosen if the dictionary only provides this spelling and simultaneously the internal attribution either has no entry or has entries for both notations. In all other but two cases the spelling without hyphen is assumed correct. The two exceptions are the cases where the internal attribution contains both spellings and the

external dictionary simultaneously provides either no or both entries. In these cases the highest number of occurrences in the internal attribution is decisive. If they are equal the spelling without hyphen is chosen.

As previously mentioned the heuristic "always use the spelling without hyphen" is the obvious way to handle hyphenation in most languages. We tested our approach against this simple heuristic with respect to the number of faulty decisions. For comparison we used a book by Guillaume Raynal in four different editions which were published in 1770, 1774, 1780 and 1820⁵ and a dictionary of the ABU : la Bibliothèque Universelle⁶ with more than 250,000 entries of common words for the second step.

The slightest relative difference between both techniques occurred in the edition of 1820 which contains 1,339 individual hyphenations of 52,372 words and 7,198 lines. Our approach resulted in 30 wrong guesses (2.240%) instead of 45 (3.368%) made by the heuristic which is a decline by the factor of 1.5. The biggest difference appeared in the edition of 1780 with 1,063 individual hyphenations, 44,078 words and 6,290 lines. While the simple heuristic resulted in 45 faulty decisions (3.814%), our approach nearly cut the number of errors in half to 24 (2.034%). Concerning the editions of 1770 and 1774 the outcome was 6 errors (0.819%) instead of 10 (1.364%), and 14 (1.317%) instead of 23 (2.164%), which is about the same level. The effectiveness of our approach becomes apparent if the four editions are considered as one text. Our approach benefits from many words with multiple occurrences in the concatenated text consisting of 155,160 words and 21,551 lines. Only 39 (1.107%) of 3,522 individual hyphenations are reversed incorrectly. In contrast the simple heuristic makes 98 faulty decisions (2.783%).

In summary our approach dominates the simple heuristic regarding the number of wrong spellings without being free of errors itself. This is important if a researcher depends on an automated disambiguation of the end-of-line hyphenated words due to the size of the text or missing expertise for deciding the correct spelling. Furthermore our approach can be helpful to considerably reduce the manual effort scholars have for checking correctness. For the tested text all but one¹ error of our approach occurred for words without any information in the internal attribution and the external dictionary. Thus a researcher can focus on these not reliable cases instead of checking every word separated by a line break. This will reduce the effort to 298 instead of 733 individual hyphenations in the edition of 1770 (40.655%), 222 instead of 1,063 in 1774 (20.884%), 197 instead of 1,180 in 1780 (16.695%), 62 instead of 1,339 in 1820 (4.630%) and 270 instead of 3,522 in the concatenated text (7.666%).

While nearly all errors of our approach occurred for words without information in both the internal attribution and the external dictionary, skipping the second step would result only in slightly increased error rates. In contrast using the external dictionary without an internal attribution would lead to apparently more errors. Both issues may be due to the historic word forms and proper names found in the text. However a large amount of entries in the internal attribution which requires that the text under consideration is relatively large seems to be the important factor for lowering the number of errors and reducing the semiautomatic effort respectively. Keeping this in mind, the approach can easily be extended by filling the internal attribution with larger corpora so that reversing the hyphenation of a relatively small text will benefit from the advantages described above. The same can be done in step two by using multiple external dictionaries.

Acknowledgments

This research was funded by the German Federal Ministry of Education and Research (BMBF) [grant number 01UG1247] as part of the project "Semi-automatische Differenzanalyse von komplexen Textvarianten" under the direction of Prof. Dr. Thomas Bremer, Prof. Dr. Paul Molitor, Dr. Jörg Ritter and Prof. Dr. Hans-Joachim Solms. Also we would like to acknowledge and thank our project collaborator Susanne Schütz.

Notes

¹ This exception is a contradictory case for the word "par-tout" in the edition of 1780 which was guessed falsely with hyphen as it was found 17 times with this spelling in the text but only without hyphen in the external dictionary. The situation is a different one if the four variants of the text are considered in all as the spelling without hyphen was used more often in the other editions than 1780.

References

1. Susan Rennie (2001) *The Electronic Scottish National Dictionary (e SND): Work in Progress* Literary and Linguistic Computing 16(2):153-160
2. Manfred Kammer (1989) *WordCruncher*: Problems of Multilingual Usage* Literary and Linguistic Computing 4(2):135-140
3. S. Hockey and J. Martin (1987) *The Oxford Concordance Program Version 2* Literary and Linguistic Computing 2(2):125-131
4. Gregory Grefenstette and Pasi Tapanainen (1994) *What is a word, What is a sentence? Problems of Tokenization* In third International Conference on Computational Lexicography (Complex'94):79-87,Budapest
5. Guillaume Thomas François Raynal *Histoire philosophique et politique des établissements et du commerce des Européens dans les deux Indes* - book six, in editions 1770 (Amsterdam), 1774 (The Hague), 1780 (Geneva) and 1820 (Paris)
6. ABU : la Bibliothèque Universelle abu.cnam.fr , retrieved 2014-02-28 10:48:28 UTC

Fractures and Cohesion: Using Systemic Functional Linguistics to Detect and Analyse Hate Speech in an Online Environment

Quinn, Deirdre
deirdre.m.quinn@nuim.ie
An Foras Feasa, NUIM

Maycock, Keith
keith.maycock@ncirl.ie
School of Computing, National College of Ireland

Keating, John
john.keating@nuim.ie
An Foras Feasa, NUIM

1. Introduction

Language acts as the lynchpin of cohesion maintaining electronic conversations across social networking sites. Analysis of that cohesion facilitates the detection of linguistic patterns that initiate and compound hate speech in online environments. This research reports on analysis of hate speech in videos and asynchronous conversation using Systemic Functional Linguistics (SFL) within social networking site. Focussing on the architecture of language and the influence of social context, SFL facilitates the analysis of language in its temporal and contextual use.¹ In applying SFL to the chosen corpus of texts the research team are building reference dictionaries of offensive words and reference catalogues for the clausal structures in which these words are used. The team is exploring the detection of and analysis of hate speech across conversation, that is across texts within social networking sites (SNS). In accumulating data that allows the expansion of dictionaries and clausal catalogues, the team is enabling the building of an automated alert system that scans texts as

they develop independently and in their engagement with other texts across time. Overall, this paper outlines the application of SFL to texts accrued from SNS that exhibit aspects of hate speech associated with dehumanisation, details the analysis of visualisations of hate speech within developing texts and demonstrates the building of an automated alert system using SFL to detect hate speech across texts.

1.1 Overview of Context

Raphael Almagor-Cohen describes hate speech as speech that intends to "injure, dehumanise, harass, intimidate, degrade and victimize the targeted groups and to ferment insensitivity against them."²

Speech acts have the capacity to carry out and compound the dehumanisation of people rendering them powerless within the confines of "fields of recognition" demarcated by the limits of language.³

Mining data retrieved from Youtube posts that dehumanise subjects provides the opportunity to analyse the overall ecology of texts as they develop online. The detection and analysis of latent dehumanisation is made possible by the application of SFL to this data thus empowering us to delineate the "fields of recognition" and the reinforcement of that field through subtle and explicit modes of language use.

1.2. Methodology

The corpus of texts used in this research are all drawn from Youtube and consist of the recording of a repeatable event in which

the videos' subjects, drug users, are subjected to dehumanising language and to an ongoing process of dehumanisation. The corpus consists of 20 videos and the associated metadata posted by unknown users of Youtube. Each video is a recording of a drug user in a public space in Ireland's capital city, Dublin. A second set of videos of similar content recorded in Glasgow, Scotland is being used by the team for comparative purposes. The team have temporarily captured the videos and compiled transcripts of the audio and of the comments posted below. Both of these transcripts and the video are treated as objects that facilitate a users' engagement with other users and with the different elements of the composite text. That is users may engage with the videos, with the asynchronous conversation that has grown in relation to the video or with media objects posted in relation to or response to these videos.

These media objects consist of mashups, memes, links to other videos and to other websites. Linguistic patterns within each type of engagement act as the creators of cohesion within the text's overall development. The team have focused on lexical cohesion as a way of analysing how items relate to each other and build the texture of the text.⁴ Lexical cohesion creates the threads through which language choice manipulates the "finite nature of language as a semiotic system."⁵ We argue that is lexical cohesion that binds nodes of conversation with other media objects and which facilitate the development of the relationship between the composite elements of the text.

By compiling dictionaries of words associated with the dehumanising aspects of hate speech the team is building a framework enabling the initial identification of dehumanisation that provides opportunities for the analysis of the "textual processes of social life".⁶ Already marginalised, the recorded drug user is drawn into the connected city as a figure of disruption who is marginalised even further by the language choices evident in the analysed transcripts. The pounding vocabulary of the audio comment adds to their marginalisation as across each video they are described in dehumanising language. This language becomes part of the rearticulation and recirculation of hate speech. SFL as a system considers language as part of a process of instantiation.⁷ That is it builds and develops texts brick by brick and interaction by interaction. In this respect, the text is considered a "complete linguistic interaction" that builds continuously.⁸ Here we

use SFL to create markers for models that will be part of the automated system that detects how hate speech builds across composite texts, between elements of the texts and across the relationships established between particular posters.

The dictionary of offensive words that dehumanise the subjects of our corpus are drawn from the encoded transcripts of both the audio and the textual comments. Markers defining lexical cohesion facilitate the exploration of the users language as it engages with the overall field of the text, with the ideational expression of the text and the text's tenor. That is markers of lexical cohesion that bind immediate nodes of conversation with each other facilitate the immediate compounding of hate speech. It also enables challenges to posts promoting dehumanising hate speech. These challenges are achieved through the fractures in language induced through interruptions in lexical cohesion. Lexical cohesion draws on the temporality of the environment in which language is used to bind conversation in immediate response structures and across more developed response structures. Interruptions to lexical cohesion can also be used to introduce new ideational expressions that counteract the binds that previously placed limits on the text's field.

Further to this, the application of sentiment analysis (SA) along with SFL methods enables the visualisation of patterns of hate speech as a text develops and as patterns gather accumulative power. Thus visualisations of hate speech in a state of emergence empower moderators to detect less explicit hate speech that may otherwise go undetected. On going analysis of 'emergent visualisations' provides further opportunities to examine the participant's use of language choice in conjunction with grammatical structures to bring cohesion to a text or to counteract the cohesion created through another user's language choices and invocation of grammatical structures. In examining the emergent visualisations the research team draws on concepts of lexical cohesion and the clausal structures of sentences to detect linguistic patterns.

2. Demo and Results

By using the concepts surrounding lexical cohesion in conjunction with the dictionary of dehumanising words the team have identified, we are able to use our custom built programme to capture and analyse conversations that have developed as part of the "complete linguistic interaction" surrounding each media object in our corpus. An example of a captured conversation is shown below in Figure 1. The media object "Dublin Junkies" is shown at the centre of the visualisation. Each node of conversation representing a cluster of chat is represented in yellow. Our custom built programme allows us to identify whether these nodes are anaphoric or exophoric conversations. That is we can identify whether the nodes are related directly to the media object or not. Anaphoric conversations, those directly related to the video "Dublin Junkies", are demarcated by the colour blue in the second image, Figure 2. Exophoric nodes, those that do not relate to the video directly, are demarcated by the colour blue. In this second image the lines joining these nodes represent the result of our SA tool. Lines joining nodes that are red denote negative sentiment and those in green denote positive sentiment. The thicker the line the more negative or positive the sentiment.

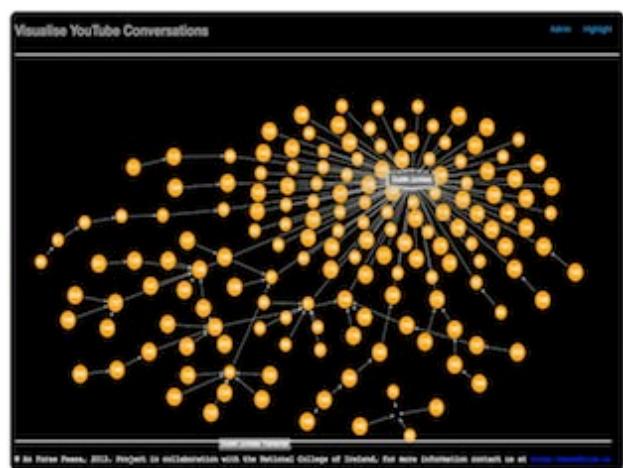


Fig. 1: Visualising the ecology of online conversation

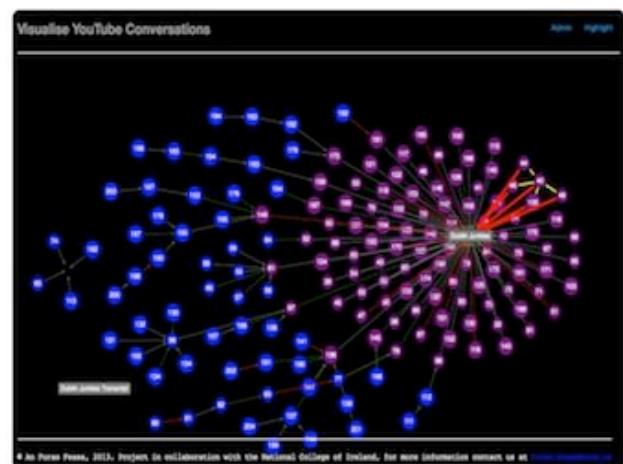


Fig. 2: Demarcating elements and relationships within developing texts

3. Conclusions

Analysis of 'emerging visualisations' points to the strong negative sentiment between exophoric nodes of conversation shown in Figure 3. The yellow lines in this figure delineate the reorganisation of the text according to lexical cohesion. This reorganisation demonstrates the capacity of both SFL and our programme to make linkages across texts as they undergo a process of instantiation.

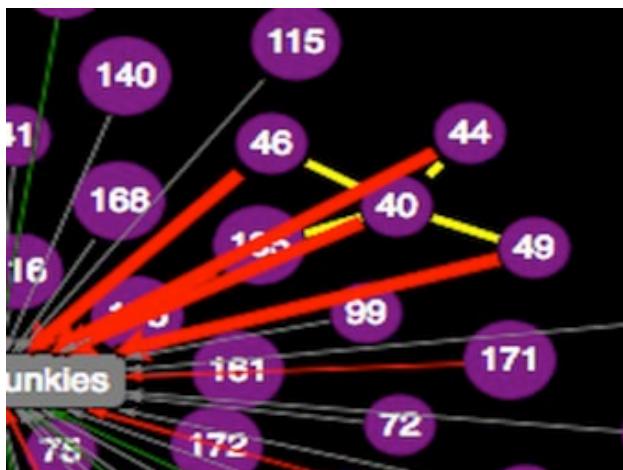


Fig. 3: Lexical Cohesion and Sentiment Analysis

Theorist Judith Butler argues that to be called a name is "to be initiated into a temporal life of language that exceeds the prior purposes that animate that call".⁹ Dehumanising language calls subjects into a disempowered temporality rendered

increasingly damaging by the capacity for rearticulation and reproduction facilitated by online environments. The use of SFL to analyse the construction of temporal fields that may be sealed by language and grammatical choices compounding dehumanisation empowers moderators to detect hate speech in an online through longitudinal analysis. The updating of visualisations also allows moderators to identify the reach of particular posters across texts enabling a vertical and an horizontal analysis of the development of hate speech online.

References

1. Almagor-Cohen, Raphael (2011), *Policing Hate and Bigotry on the Internet* in Policy & Internet 3.3 7.
2. Butler, Judith (1997), *Excitable Speech: A Politics of the Performative* (London: Routledge) 7.
3. de los Angeles Gomez Gonzalez, Maria (2011), *Lexical Cohesion in Multiparty Conversations* in Language Sciences 33 168.
4. Eggins, Suzanne (2004), *An Introduction to Systemic Functional Linguistics* (2nd Edition) (London: Continuum) 14
5. Eggins, Suzanne (2004), *An Introduction to Systemic Functional Linguistics* (2nd Edition) (London: Continuum) 2.
6. Halliday, M.A. K and Jonathan J. Webster (2009) (Eds) *Continuum Companion to Systemic Functional Linguistics* (London: Continuum International) 4.
7. Eggins, Suzanne (2004), *An Introduction to Systemic Functional Linguistics* (2nd Edition) (London: Continuum) .
8. Butler, Judith (1997), *Excitable Speech: A Politics of the Performative* (London: Routledge) 14.

Framework for Quantitative Analysis of Scripts

Rajan, Vinodh

vrs3@st-andrews.ac.uk
University of St Andrews

1. Introduction

1.1 Overview

Scripts are usually seen as simple carriers of languages. Research on scripts until recently has been minimal and niche, except for the field of paleography. Scripts are an important part of the cultural heritage of humanity and its analysis and study requires more research. Fortunately, there is a growing interest in analysis of scripts. Altmann ¹ published a volume titled "Analyses of Scripts: Properties of Characters and Writing Systems" to explore various properties of writing systems and scripts such as complexity, ornamentality and distinctivity.

Changizi et al ² discuss the various contour configurations of written symbols and their similarity to the environment in which they were produced. They also study the distribution of the configurations of various scripts. They ³ further discuss the character complexity and the redundancy of stroke combinations of various writing systems in human history. It is to be noted that analysis in [2] [3] and most methods in [1] were performed manually. Traditionally, analysis and study in paleography have also been done manually. Digital paleographic methods are at present making more inroads into the field. However, applying quantitative analysis on paleographic data is not yet popular and standardized ⁴. This is partially due to the difficulty of quantifying paleographical features, and partially due to the lack of defined metrics with theoretical and qualitative underpinnings.

1.2 Proposed Framework

We propose here a quantitative analysis framework for scripts that is largely computational and requires minimal user interaction. We do not particularly aim at providing a completely autonomous framework, but rather to aid the user as much as possible, with the ability to manually intervene/override as required. The framework is based on various methods and techniques employed in the field of graphonomics. Our framework is grounded on the principles of handwriting production and handwriting analysis.

We also explore various features used in the related area of gesture design and recognition and its application to the analysis of scripts. Through this, we attempt to find relevant metrics (with qualitative significance) with sufficient evaluation that might be used for glyphs and scripts for various purposes such as classification, visualization etc. The computed quantitative features could serve as descriptors for scripts, and be used for comparing and analyzing scripts. This is especially applicable to the field of paleography, where such quantitative features are much needed.

2. Quantitative Analysis Framework

Our proposed framework consists of the following modules.

2.1 Spline Conversion

The characters of scripts are externally represented as B-splines. B-splines are very efficient in preserving the shape and curvature of glyptic segments. Additionally, they can be manipulated without significant effort. Rather than representing glyphs as pixelated data, converting them into splines eases analysis. This conversion of the glyptic shape of a character can be done automatically or manually. In a manual process, the user defines each shape of a character directly using a set of B-splines, or explicitly draws the shape, which is then internally converted into B-splines. An automatic conversion of glyphs involves thinning and then its conversion into splines.

2.2 Trajectory Reconstruction

The shape of a character is static and does not contain all information required for analysis. Dynamic information relating to pen movements are not present in the shape. Trajectory reconstruction attempts to recover this temporal information ⁵. This kinematic information is essential in defining the character. With paleographic scripts reconstructing dynamic information is necessary as the trajectory is usually unknown. Also by altering the trajectory, the changes in dynamic features can be observed.

The recovery is performed by conducting a global search using a set of heuristics, such as length minimization and curvature minimization ⁶. Especially in case of paleographic scripts, the algorithm is able to provide several alternative viable writing trajectories.

2.3 Stroke Segmentation

Characters are best analyzed as sequences of natural strokes. Breaking them down into basic strokes is the optimal way for analyzing written characters. It also enables us to understand the process of handwriting. Stroke segmentation retrieves the structure of a character based on its trajectory. This is performed by segmentation of the character at various important landmark points of the recovered trajectory such as the extrema of curvature ⁷.

2.4 Character Representation

Writing is usually considered to consist of two fundamental types of strokes – up-strokes and down-strokes ⁸. This distinction is necessary since both these two types behave

differently. It has been proved that down-strokes usually do not show a lot of variation in handwriting compared to up-strokes⁹.

A character is internally represented as a set of strokes. This is consistent with the way that the character is internalized and produced by humans. This allows us to derive better features that are more natural and descriptive. Later, it will be possible to apply handwriting modelling to generate alternative scribal variants.

2.5 Feature Extraction

For quantitative analysis, features need to be computed from the characters. These serve to quantify several aspects of the characters. We considered various features used in the field of gesture recognition^{10 11 12 13} and found the features listed below to be relevant to the analysis of characters. We also propose some features in addition to those found in the literature. These features/metrics could additionally serve as descriptors for the scripts. As quantitative features these can be widely used in analysis and/or visualization.

2.5.1 Production Features

The effort that is required to realize and produce a character is an important element of its analysis. It is related to the *number of velocity inversions*, *number of velocity breaks*, *number of pen lifts*, etc., which are computed from the stroke representation of the character. These features are calculated from the temporal information that was reconstructed at the earlier stage. These are relevant as they quantify the dynamic handwriting behavior present in the character.

2.5.2 Geometric Features

Geometric features throw more light on the visual aspect of characters. These are essential for the study of the judged (visual) complexity¹⁴. The features that relate to visual appearance are *compactness*, *openness*, *number of crossings*, *average curvature*, *sum of internal angles*, *bounding area*, etc.

Some of these, like *compactness* and *openness*, are ratios of several parameters such as *length of strokes*, *distance between first and last points*, while others, like *average curvature*, are derived directly from the glyph structure.

2.5.3 Cognitive Features

Though cognitive features cannot be directly measured, some cognitive features could be interpreted from the geometry of a character. The number of *unique landmark points* required to plan the trajectory is a possible feature that has correlation with the cognitive load of the glyph. An additional measure is the *number of minimal points* required to recreate the character.

2.5.4 Stroke Features

Stroke based features such as the *primary direction* of the glyph, *ratio of upstrokes to downstrokes*, *direction change*, *histogram of inter-stroke angles* etc. are also computed for a character.

3. Prototype Implementation

A prototype of the framework has been implemented in Python with the modules discussed above. We are planning to analyze the development of Indic scripts using the framework. The source code will be released under an open source license when the project reaches maturity. Its repository will also include a complete set of Indian paleographic scripts. Below, we briefly describe functionality yet to be implemented in the prototype.

From the perspective of paleographic scripts, the available glyphs in the literature are usually very noisy. Scanning and importing them would require several layers of pre-processing and noise-removal. For modern scripts, importing the Bezier curves from the respective fonts could be done directly.

Trajectory reconstruction has been implemented only for single stroke characters. A primitive implementation exists for multi-stroke characters. This needs to be made more rigorous and accurate.

Based on the stroke structure of a glyph several additional features such as *entropy of writing* could be calculated as required. Also, the features are to be used for normalized glyph shapes. The behavior of the features with respect to various scribal variants needs to be analyzed further.

Proper visualization of the various quantitative features would help to better study and understand the characters within a script and also compare several scripts. Such visualization is particularly helpful in studying paleographic scripts and analyzing the changes that took place over time. Various statistical analyses of the features and visualization techniques would be built into the implementation.

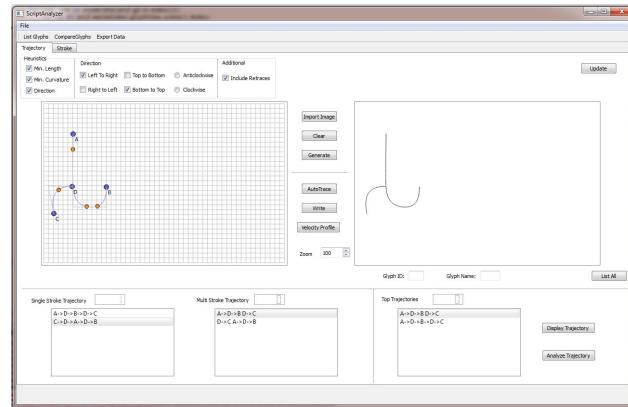


Fig. 1: Spline Conversion and Trajectory Generation

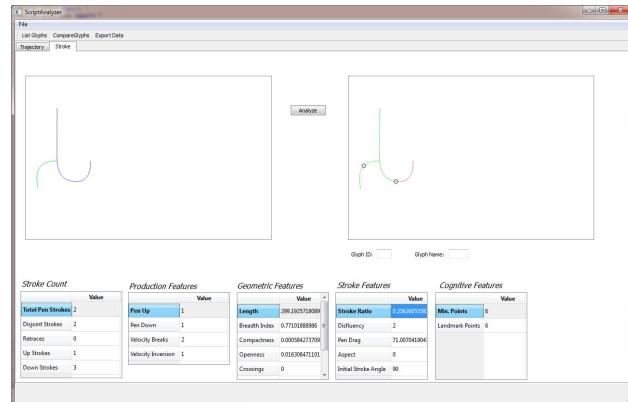


Fig. 2: Stroke Segmentation & Feature Extraction

4. Similar Projects

There are other digital paleographical projects, such as Hand Analyser by Peter Strokes, which work on pixelated images. The current project is more focussed on using *strokes* to derive various features. Integration/Adaptation of those techniques needs to be further looked into.

5. Summary

We have presented a computational framework for quantitative analysis of scripts. The framework requires minimal user interaction, and is based on the principles of handwriting analysis and handwriting production. We also present a prototype implementation of the proposed framework. We believe this framework and its implementation would facilitate more quantitative study on scripts.

References

1. Altmann, Gabriel, and Fan Fengxiang (2008), eds. *Analyses of script: properties of characters and writing systems*. Vol. 63. Walter de Gruyter. APA.
2. Changizi, Mark A., et al. (2006) *The structures of letters and symbols throughout human history are selected to match those found in objects in natural scenes*. The American Naturalist 167.5: E117-E139.
3. Changizi, Mark A., and Shinsuke Shimojo. (2005) *Character complexity and redundancy in writing systems over human history*. Proceedings of the Royal Society B: Biological Sciences 272.1560: 267-275.
4. Stokes, Peter. (2009) *Computer-aided palaeography, present and future*. Kodikologie und Paläographie im digitalen Zeitalter - Codicology and Palaeography in the Digital Age. Schriften des Instituts für Dokumentologie und Editorik, 2: 309-338.
5. Doermann, David S., and Azriel Rosenfeld. (1995) *Recovery of temporal information from static images of handwriting*. International Journal of Computer Vision 15.1-2: 143-164.
6. Jager, Stefan (1996). *Recovering writing traces in off-line handwriting recognition: Using a global optimization technique*. Pattern Recognition, , Proceedings of the 13th International Conference on. Vol. 3. IEEE, 1996.
7. Li, Xiaolin, Marc Parizeau, and Réjean Plamondon (1998). *Segmentation and reconstruction of on-line handwritten scripts*. Pattern recognition 31.6: 675-684.
8. Noordzij, Gerrit, and Peter Enneson (2006). *The stroke: Theory of writing*. Hyphen.
9. Teulings, Hans-Leo, and Lambert RB Schomaker (1993). *Invariant properties between stroke features in handwriting*. Acta psychologica 82.1: 69-88.
10. Rubine, Dean (1991). *Specifying gestures by example*. Vol. 25. No. 4. ACM.
11. Long Jr, A. Chris, et al (2000). *Visual similarity of pen gestures*. Proceedings of the SIGCHI conference on Human factors in computing systems. ACM.
12. Willems, Don, et al (2009). *Iconic and multi-stroke gesture recognition*. Pattern Recognition 42.12: 3303-3312.
13. Tucha, Oliver, Lara Tucha, and Klaus W. Lange. (2008) *Graphonomics, automaticity and handwriting assessment*. Literacy 42.3: 145-155.
14. Stenson, Herbert H. (1966) *The physical factor structure of random forms and their judged complexity*. Perception & Psychophysics 1.9: 303-310.

Macro-Etymological Textual Analysis

Reeve, Jonathan Pearce

reeve@nyu.edu
New York University, United States of America

The English language has continually borrowed from foreign languages—close to 30% of modern English words are loanwords from French, and another 30% are borrowed from Latin. These words are often concentrated in semantic frames associated with their origin languages—legal vocabulary contains a preponderance of words of French origin, and the vocabulary of the natural sciences contains many words of Latin and Greek origin. The etymology of words in a text, therefore, may be suggestive of its context or its level of discourse. Should a writer choose the Latinate term “masticate” over the Anglo-Saxon term “chew,” for instance, one might assume a scientific context or a high level of discursive formality. By computing the proportion of origin languages for all the words of a given text, we may quantify stylistic properties that are potentially revealing about the text and its rhetorical modes.

The Macro-Etymological Analyzer is a computer program that I wrote for this purpose. Written in PHP on a LAMP stack, it is a web app accessible at <http://jonreeve.com/etym>, and is freely

available for all to use, modify, and distribute under the GPLv3. It accepts as input a user-uploaded text file, and looks up each word in Gerard de Melo’s Etymological Wordnet database. These words are then counted by language of origin using two generations of language ancestry, and then categorized by language family. The results are displayed as a pie chart made with the Google Data Visualization API, along with a CSV log file which can be used for comparative analyses. Currently, the program accepts only English texts, but the database supports queries from any source language, and plans are in place to make the program fully multilingual.

Figure 1 shows the proportions of Latinate words—words descended from Latin or romance languages—for each of the 15 genres in the Brown Corpus. Learned texts and government documents show the highest proportions of Latinate words, whereas romance and adventure stories show the lowest. The same textual categories sorted by proportion of Hellenic words (words of ancient Greek origin) show changes in certain categories—religious language exhibits a higher rank, and that of mystery stories is ranked lower than in the Latinate scale. These data suggest that a high proportion of Hellenic words is correlated with religious language, among other genres, and that a high proportion of Latinate words is correlated with learned language. Once literary works are analyzed with this method, these hypothetical correlations become potentially useful as literary critical tools.

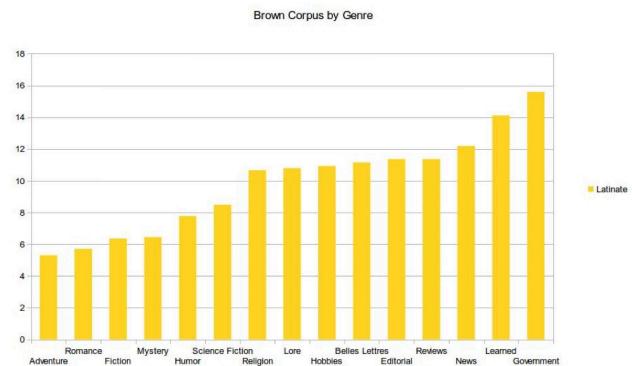


Fig. 1: Brown Corpus Genres

In one such analysis, the chapters of *A Portrait of the Artist as a Young Man* were run through the Macro-Etymological Analyzer. This novel, James Joyce’s Bildungsroman, is known for its style—one that mimics each progressive age of its protagonist Stephen Dedalus. Early chapters, when he is young, are written with infantile language; later chapters are written with more elevated language. The program’s results quantify this stylistic mode, to some degree—Chapters 1 and 2 show low proportions of Latinate words, whereas later chapters show higher proportions, as shown here in Figure 2. The fact that the proportion of Latinate words begins to plateau starting with Chapter 3 might be used to argue that Stephen has at this young age already reached a precocious maturity of vocabulary, which may reflect his study of Latin.

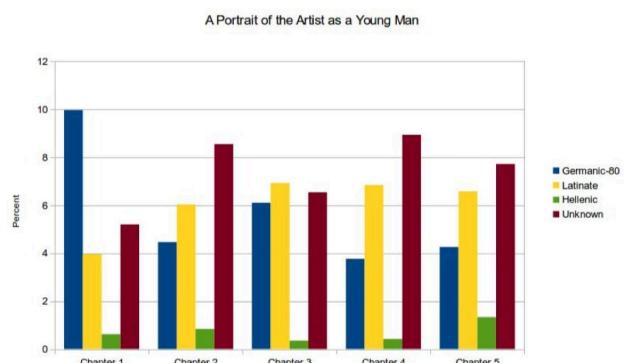


Fig. 2: A Portrait of the Artist as a Young Man

In another analysis, the extracted monologues of the seven narrators of Virginia Woolf's novel *The Waves* were computed with this program. As shown in Figure 3, the two university-educated characters, Bernard and Neville, show the highest proportions of Latinate words, while the housewife Susan shows the lowest. In fact, the male characters rank higher for Latinate words than the female characters—this would be an interesting starting-point for a discussion of gender in *The Waves*, especially framed by Woolf's much-discussed writings on gender politics.

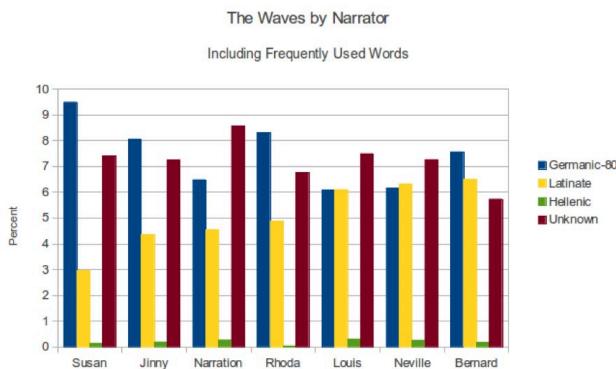


Fig. 3: The Waves Narrators

The Macro-Etymological Analyzer was also used to chart variations between editions of a text. The seven revisions of Whitman's *Leaves of Grass* made available by the Whitman Archive were analyzed with this program. The results show a gradual increase in Latinate words from the 1855 edition to that of 1891-2. This might be used to argue that Whitman inflated his style with each revision, introduced foreign loanwords as he gained a more international reputation, or used a greater breadth of words as his vocabulary increased.

These experiments were not without their surprises, of course. An early test of selected books of the King James Bible seemed promising, as it revealed the gospels Matthew, Mark, Luke, and John to have much higher proportions of Hellenic words than other books (see Figure 4). Unlike the books of the Old Testament, which were mostly written in Hebrew, these books were translated from the Greek—a fact which might seem to explain the presence of Hellenic words. Upon closer examination, however, the program was discovered to be counting the etymology of frequently-mentioned names like “Jesus” among words of Hellenic origin, and it was these names that accounted for most of the Hellenic words. Although the language of the source text did not prove to be the determinant here, this discovery may yet be valuable for other reasons—the synoptic gospels of Matthew, Mark, and Luke show similar portions of Hellenic words, whereas that of John is 100% greater. This would seem to support the hypothesis that the synoptic gospels were adapted from a common source text, whereas that of John had an independent source.

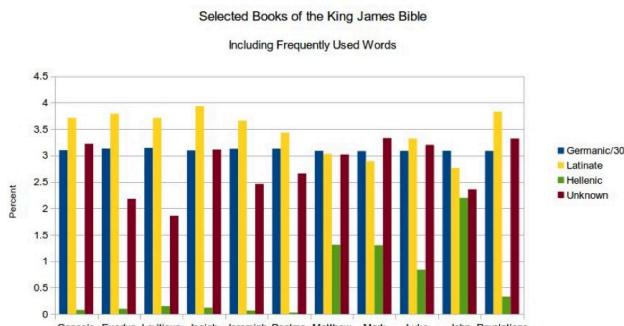


Fig. 4: KJV Bible

A number of other experiments were also conducted, and are described in this paper. Included among texts analyzed by the Macro-Etymological Analyzer were: selected Canterbury tales (in modern English translation), a series of early and late

Henry James novels, a collection of Victorian novels compared with a collection of modernist novels, and groups of French and German novels in English translation. Questions to be explored include:

- Do translated works show a larger-than-normal proportion of words with etymological origins in the language of the source text?
- Given a large enough data set, can linguistic trends (such as a general decrease in the use of Latinate words) be detected with this program? Can macro-historical events such as the Scientific Revolution be detected?
- Do male and female writers of the 19th century differ in the origin-types of words they use?
- Can the semantic frames in which these etymological groups of words are concentrated be explained historically, such as through the habits of the French-speaking English aristocracy in the era following the Norman Conquest?

Finally, this paper will discuss how this new tool might contribute to the suite of computational stylistics tools already available, and how macro-etymology might constitute a new metric that could be used towards stylistic fingerprinting or authorial detection.

Crowdsourcing Performing Arts History with NYPL's ENSEMBLE

Reside, Doug

dougreside@gmail.com

NYPL

1. Introduction

The New York Public Library for the Performing Arts holds in its collection over one million programs documenting a large number of the major theater, music, and dance events performed around the country since the end of the Civil War. Although the collection grows each month, the Library estimates that it currently holds approximately 125,000 dance, 400,000 music, and over one million theater programs. These programs are valuable as individual artifacts, of course, but as an aggregated collection they serve as a sort of analog database of performing arts history. Unfortunately, querying this “database” is, at present, very inefficient. The materials are available only to researchers who come to New York to view them in person, and can only be viewed one at a time. Further, many are printed on crumbling paper that may not survive many more examinations by even careful researchers.

In early 2013, motivated both by our responsibility to preserve these artifacts and out of a desire to better expose the data they contain, we launched an effort to create digital images of our program collection and organize a crowd-sourced effort to transcribe and structure the information contained within it.

The project, launched in beta under the name *Ensemble*¹ in June of 2013, is now part of a new NEH-funded Digital Humanities Implementation grant to create tools for crowd-sourced transcription projects². This paper will discuss the lessons the team learned from the beta release as well as the modifications we are planning for the upcoming full release in 2014.

2. Behind the Beta

Although it is our goal is to scan and transcribe every program in the Library’s collection, for the beta release we scanned 5 reels of microfilm containing 200 programs connected to theatrical productions in New York City performed between 1860 and 1930. We selected this content for several reasons:

- The relatively low cost and high efficiency of microfilm scanning allowed us to add a relatively large number of programs to our initial set very cheaply and quickly. Although in most cases we would prefer to digitize originals, the programs preserved on these reels no longer exist in our collections.
- Performing arts events from this period are not well-documented by other online databases (such as the Internet Broadway Database³ or Playbill Vault⁴).
- These programs are almost certainly in the public domain; therefore they can be scanned and published online in their entirety. (If programs printed between 1923 and 1950 ever were in copyright, they were not likely not renewed and so passed into the public domain 28 years after publication).
- This period was an especially fertile time in the development of the American performing arts; Carnegie Hall was built⁵, and Vaudeville and American Musical Theater both developed during these decades.

3. What we learned

The beta release of *Ensemble* has, as of this writing, produced almost 11,500 transcriptions of data from our initial test set of 200 programs. Although this is significant, it falls far short of the activity seen by other crowd-sourcing projects released by NYPL Labs. In its first three years, the menus transcription project has had over 1.2 million dishes transcribed. Over 60,000 buildings were checked in the first days of the “Building Inspector” app⁶. By comparison, participation in *Ensemble* is very low.

In part, these lower numbers may reflect the relative difficulty of the task *Ensemble* assigns to its users. Rather than asking for a simple transcription (as the Menus project does), users of *Ensemble* are required to identify relationships among text on the page (and occasionally to bring to it their own understanding of the theater industry). For instance, a program in our collection purports to be a record of “Jesse L. Lasky’s Aristic Novelty: Fleurette.” A user assigned this program must determine whether Jesse L. Lasky is the playwright, the producer, or perhaps the director.

In some cases, our interface may not even have an appropriate category. Lacking a consistently adopted schema for performing arts data, the user is required to engage in a bit of amateur taxonomy. In our first official release, we plan to revise and publish our schema, and make it easier for novice users to perform less demanding tasks while saving more challenging assignments for “advanced” levels of the game. Zooniverse’s transcription project, *Old Weather*⁷, has had success with a similar approach.

Following the model of the citizen sciences like Zooniverse’s Galaxy Zoo, *Ensemble* requires “agreement” by several users before accepting a crowd-sourced transcription as correct. The level of agreement among different transcriptions of the same text is processed by our systems and will eventually be used to determine what assertions are stored in the database that users of *Ensemble* construct. In our initial version, we attempted to expose the quality assurance/“user-agreement.” Our hope was that those who were suspicious of the accuracy of any database constructed by the “crowd” would be somewhat reassured after they understood how the process worked. More often, though, we found that users who were the first to transcribe a fact, and then saw that the system had a low “degree of confidence” in the work they had just submitted (since no one else had yet “agreed” with them) misunderstood what they were being told and felt either insulted or disheartened. We quickly removed these visualization (although we may want to find a way to incorporate them in a more clearly contextualized way in the final version of the tool).

4. Potential uses of the data

Of course, the reason for engaging the crowd to produce this dataset in the first place is based on the assumption that it will be useful to future researchers. Some initial use cases have imagined include:

Aggregating archives: At present, researching historic performing arts events can be difficult as most of the primary sources are held in collections centered on a particular person. For example, if a scholar is researching the George M. Cohen musical *Little Johnny Jones* he or she will quickly discover that there is no large Little Johnny Jones collection at any major library. Once all of the data in our programs is available, however, a researcher could write a computer program that, given a title of a show, could generate a list of people associated with it and automatically search Worldcat for libraries that hold archives related to these people.

Discovering untold biographies: The lives of star performers and successful writers are often studied, but the careers of those members of a production whose role is less visible, but no less vital, often go unchronicled. *Ensemble* will enable researchers to track, for instance, which stage managers are most often associated with successful plays, which celloists were featured in the best orchestras of the 1920s, and, which constellation of artists and technicians is most often associated with the success or failure of the production of a Shakespeare play.

Mapping the arts: Where in New York City in 1920 would one mostly likely find an opera performed? What about a burlesque? A jazz concert? By opening up the data in the programs, it will be possible for software developers and geographers to combine the performance data with our digitized historical map collection and plot the kinds of performing arts events performed in particular regions of the City during a defined time period. This data may confirm or overturn scholarly assumptions about the geographical history of the city.

It is possible that the most illuminating and exciting uses of the data in these programs have yet to be imagined because, at the moment, surprisingly little of this information from this period is available at all. It is our hope that *Ensemble* will soon become the backbone of an extensive, linked, open set of performing arts data that will allow researchers of all kinds to discover new information about the rich history of the performing arts in New York.

References

1. **Ensemble** (2013). *Ensemble: Help Build an Open Database of the Performing Arts*. ensemble.nypl.org/ Web. 27 Oct. 2013.
2. *Announcing 6 Digital Humanities Implementation Grant Awards (July 2013) | National Endowment for the Humanities* (2013). www.neh.gov/divisions/odh/grant-news/announcing-6-digital-humanities-implementation-grant-awards-july-2013 Web. 27 Oct. 2013.
3. *IBDB: The Official Source for Broadway Information*. ibdb.com Web. 27 Oct. 2013.
4. *The Largest Internet Database of Broadway Information - Playbill Vault*. playbillvault.com Web. 27 Oct. 2013.
5. *History of the Hall | Carnegie Hall*. www.carnegiehall.org/ History Web. 27 Oct. 2013.
6. *Building Inspector by NYPL Labs*. buildinginspector.nypl.org Web. 27 Oct. 2013.
7. *Alexandra Eveleigh, Charlene Jennett, Stuart Lynn and Anna Cox* (2013). *I want to be a Captain! I want to be a Captain!*: Gamification in the Old Weather Citizen Science Project. uwaterloo.ca/gamification/sites/ca.gamification/files/uploads/files 26 Oct 2013
8. *Galaxy Zoo*. www.galaxyzoo.org Web. 27 Oct. 2013.

Two new tools for multimodal editions

Reside, Doug
dougreside@gmail.com
NYPL

1. Introduction

From the earliest e-Text centers at Oxford¹ and the University of Virginia², through the development of the TEI in the late 1980s³, to the publication of image-based editions in projects like The William Blake Archive⁴, digital editing projects have long been a core activity of the Digital Humanities. Until very recently, though, the limitations of reliably available technology and complicated intellectual property laws have kept all but the most adventurous editors from venturing beyond the media of text and image to multimodal editions that incorporated video and audio recordings. Over the past year, however, the New York Public Library for the Performing Arts has released two new digital editing projects: an online tool for producing editions of dance video and a mobile app for publishing multimodal, variorum editions of musical theater libretti. This paper will examine both projects and what they imply for the future of critical edition building in the 21st century.

2. Antecedents

Although most digital editions produced by digital humanities scholars have presented only text and image, over the past two decades, a few pioneers attempted to produce multimodal texts of various kinds. In the 1990s the Voyager Company produced a series of Hypercard and CD-ROM based multimodal editions (such as a Companion to Beethoven's 9th Symphony)⁵. A 2002 project led by Janet Murray produced a critical digital edition of the film, *Casablanca*⁶. More recently, the Maryland Institute for Technology in the Humanities, Music Theatre Online, produced with an NEH Digital Humanities Startup grant by the Maryland Institute for Technology in the Humanities, linked several libretti of the 2008 musical Glory Days to mp3 files from live performances⁷. Although each was an interesting experiment that advanced the field of digital editing, the Voyager CD-ROMs were never financially successful⁸, copyright restrictions prevented the widespread release of Murray's Casablanca edition, and Music Theater Online never found a large audience.

Over the last five years, though, several new technologies have emerged which have made the production and dissemination of multimodal editions much easier. The now widely adopted, video-and-audio-friendly HTML 5 specification and code libraries like Popcorn.js have made it much easier to embed and control audio and moving image recordings on web pages without relying on external plugins like Adobe's Flash or Microsoft's Shockwave. Cloud-based streaming services with public APIs such as Rdio⁹ and YouTube¹⁰ allow digital editors to make use of content hosted by third parties (making rights clearance less of a concern). Further, the ever accelerating migration of users from desktops and laptops to mobile devices and their app-based ecosystem makes it possible to publish these editions on hardware that has become a comfortable technology for long-form reading (thereby overcoming some of the ergonomic obstacles that prevented wide-spread adoption of many earlier digital editions).

3. Our projects

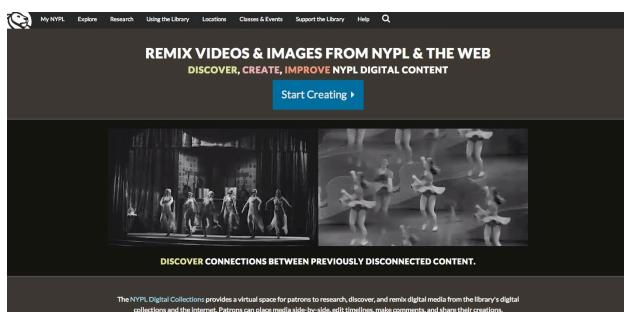


Fig. 1: Two video synchronized and shown together



Fig. 2: The editing environment

In October of 2013, the New York Public Library for the Performing Arts released a collection of over 1000 hours of video of dance performance via a new web-based collections portal. In addition, we released a web-based video editor, based on the JavaScript library Popcorn.js, which allows users to synchronize different videos of the same event (or work) together to be watched side-by-side (thereby creating a kind of multimodal variorum edition) [see figure A]. The tool also allows segments of multiple videos from various sources (e.g. NYPL Archives and YouTube) to be edited together into a playlist and annotated (either with text or additional videos) to create a video critical edition [see figure B]. The individual elements of any "edition" created with this tool will only be played in locations where rights agreements allow (for instance, in one of the 88 branches of New York Public Library), but any publicly viewable content and textual annotations can be seen anywhere in the world.

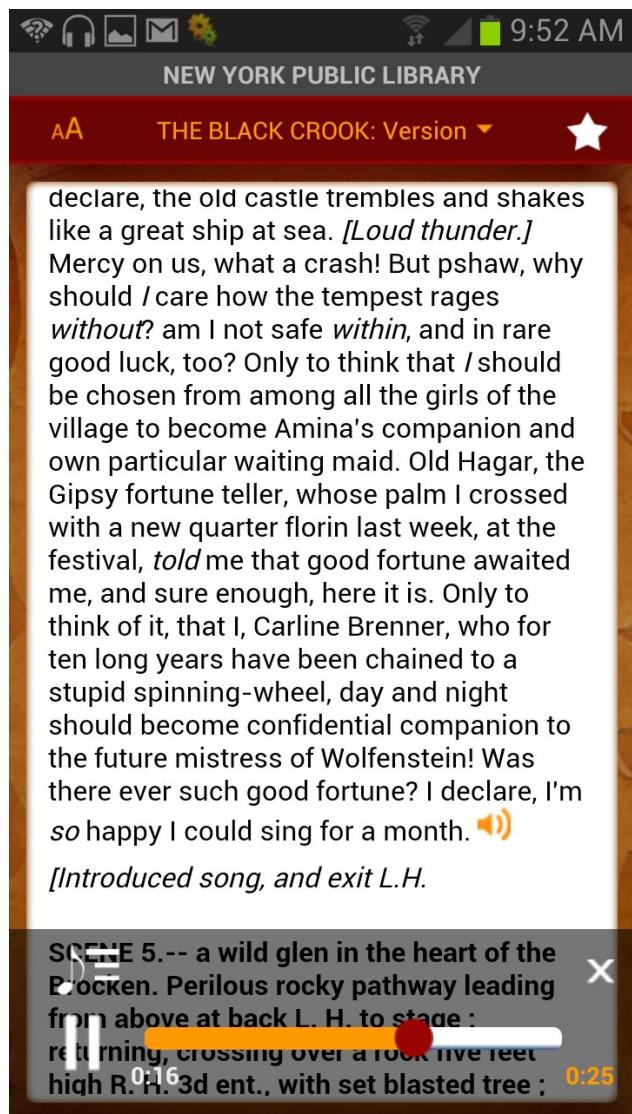


Fig. 3: Libretto: The Android App

In February of 2014, the Library will also release an NEH-funded eBook app (provisionally titled Libretto) for Android operating systems capable of presenting variorum editions of texts linked to audio recordings. Readers will be able to switch among variant versions of the same text, and (through an implementation of the new ePub 3.0 ebook standard) click portions of the text to hear associated audio (e.g. music associated with lyrics in an opera libretto) [see figure C]. In most cases, the music should be bundled with the text, but, in cases where obtaining a license to redistribute the music proves impossible, the editor may link the reader to an online store (such as Amazon.com) to purchase recordings which will automatically be synched to the text based on timestamps included in the editorial metadata.

4. Analysis

It is our hypothesis that scholarly users of the multimodal editions created for these environments will find the multichannel approach we have taken useful. However, relatively little research has been done in exactly this context. The Voyager company's CD-ROMS and the Casablanca edition were similar, but designed only for desktops and optical media, and so represented a different experience than we hope to provide with both our mobile eBook reader and the web-based video editor. However, by July, the video editing tool will have been public for nine months and Android app will have been public for about five. During this period we will track usage of both, and in this short paper will report on the response of both editors and "readers" of these new tools for multimodal editions and what it suggests for the scholarly editors of the future.

References

1. **IT Services, 13 Banbury Road.** *The University of Oxford Text Archive.* Text. ota.ahds.ac.uk/Web . 28 Oct. 2013.
2. **About The Etext Center | University of Virginia Library Digital Curation Services.** www.digitalcurationservices.org/digital-stewardship-services/etext Web. 28 Oct. 2013.
3. **TEI guidelines.** www.tei-c.org/Guidelines/P5/ .
4. **The William Blake Archive Homepage.** www.blakearchive.org/blake/ Web. 28 Oct. 2013.
5. **Brown, Geoffrey.** *Developing Virtual CD-ROM Collections: The Voyager Company Publications.* International Journal of Digital Curation 7.2 (2012): 3–20. www.ijdc.net. Web. 28 Oct. 2013.
6. **Here's Looking at Casablanca.** National Endowment for the Humanities. Accessed October 28, 2013. www.neh.gov/humanities/2005/septemberoctober/feature/heres%2080%99s-looking-casablanca .
7. **Music Theatre Online.** mith.umd.edu/mto Web. 28 Oct. 2013.
8. **Reid, Calvin.** *Voyager Shakeup: In Wake of Stein's Exit, It's up for Grabs.* Publishers Weekly, November 18, 1996. General OneFile.
9. **Rdio - Welcome to the Rdio API.** Accessed October 28, 2013. developer.rdio.com .
10. **YouTube JavaScript Player API Reference - YouTube — Google Developers.** Accessed October 28, 2013. developers.google.com/youtube/js_api_reference .

Using Computer Vision to Improve Image Metadata

Reside, Doug
dougreside@gmail.com
NYPL

Introduction

We are approaching a golden age in the study of visual art and photographs. Many museums, libraries, and universities have digitized large portions of their collections and have made the images, and associated metadata, available for study. This process has been a major boon to art historians, collectors, and other researchers. Instead of calling individual items one at a time in library reading rooms or digging through old, expensive, incomplete, and often out-of-print art catalogs, a researcher can simply type a query into a library or museum database and retrieve large sets of images. However, while the rate of digitization has lately increased, this has, at times, come at the expense of detailed cataloging. Even before the era of mass digitization, catalogers struggled to identify certain works of art (sometimes due to a lack of collaboration between institutions). Today, simple digitization is often faster and cheaper than expert cataloging, and so many works of art and photos appear in repositories with limited metadata using inconsistent schema or vocabularies¹. The result is that while there are more works of art online than ever, it is still difficult for researchers to find the images they seek. However, advances in automated fuzzy image recognition may allow researchers to discover relevant content even when available metadata is limited. In this paper, we will examine two test cases--photographs of theatrical performance with unidentified actors, and mis-attributed Japanese woodblock prints--to demonstrate how image search algorithms can be used both to locate content and help researchers understand it better.

Overview of the test cases

Woodblock prints

In December of 2012, John Resig (Visiting Researcher at Ritsumeikan University and creator of the jQuery JavaScript library) released Ukiyo-e.org: a database of Japanese woodblock print images with metadata harvested by traversing the publicly-accessible digitized collections of prints at the targeted institutions. The images were copied and saved to a separate server for faster access (a technique which avoids overburdening the institutions by loading the images directly from their websites). The information on the website is organized broadly by artist and time period on the homepage, but is primarily designed to be used as a search engine allowing users to search both by text and by images. The database currently contains over 213,000 prints from 24 institutions collected from late 2011 to late 2012.

One of the most important features of the Ukiyo-e.org² website is its ability to do real-time analysis on the images it holds for comparison and searching. There is frequently disagreement among major institutions regarding the attribution, dating, titles, and other information associated with a print. Because of this incongruous metadata, it becomes virtually impossible to find similar prints among multiple institutions. The one piece of information that is never under contention, however, is the image of the print itself. The image that is presented by most institutions usually includes a full, straight-on photograph of the print (or prints, if it's a diptych, triptych, or similar). By ignoring the metadata provided by the institutions and comparing only the actual contents of the images, it is possible to find similar-looking prints at different institutions.

Theater Photographs

Along with works of art, many libraries and archives have recently begun to publish large sets of photographs, often depicting unidentified people. In some cases, the lack of any additional information makes identification of the faces in the photographs all but impossible. However, images from performing arts collections often feature well-documented events with widely recognizable people appearing alongside lesser known figures. A rehearsal shot of a musical comedy, for instance, might feature a star in front of a chorus of anonymous extras. Metadata for such photographs may identify the star,

and the title of the piece in which he or she is performing if it is known, but the supporting cast is generally left unidentified.

In early 2012, Doug Reside, Digital Curator for the Performing Arts at New York Public Library, began a series of experiments to attempt to identify these performers. Over 90,000 theater photographs are now available on the Library's website with varying degrees of metadata³. In most cases, the work being presented is identified. Information about the cast and crew of these productions often can be harvested from other online databases such as Playbill Vault⁴, the Internet Broadway Database⁵, and DBpedia⁶. Given infinite time, a human investigator could theoretically identify many of the anonymous people in the Library's photographs by finding all instances of the face in any online photograph, and then using additional datasets to determine the most likely name associated with it. An otherwise anonymous person, might, for instance, be identified in a newspaper photograph or in a headshot in theater program. Similarly, it might be possible to identify an otherwise anonymous actress if the shows in which her face appears uniquely match her resume as constructed from published cast lists.

Methodology

Both test cases would benefit from computer vision algorithms capable of searching a corpus of images for a set of very similar (but not necessarily identical) images. Although research in this area began decades ago, implementations capable of comparing thousands to millions of images from various sources simultaneously have only emerged very recently⁷. The general availability of this technology, however, has been mixed. Tools such as imgSeek⁸ have made rudimentary image comparison technologies available for use in Open Source projects, but at present commercially-available tools with public APIs (such as TinEye's MatchEngine⁹) provide faster image analysis with a greater level of clarity. Neither tool, however, is exactly suited for facial recognition, which requires the ability to identify a face pictured at different angles, under different lighting, in front of varying backgrounds, and at varying sizes (depending on the distance of the subjects for the camera).

The MatchEngine tool, while a commercial service, is well suited for finding images that are close matches of one another, or even partial matches embedded inside a larger image (as in the case of triptychs). Like imgSeek, MatchEngine was able to find images by upload and quickly process newly-added images. Resig tested both MatchEngine and imgSeek during the development of Ukiyo-e.org and found that MatchEngine was much better at finding exact matches, ignoring differences in color, and finding prints (or portions of prints) inside other print images.

With an effective image similarity engine it became possible to develop many new tools to aid woodblock print researchers. Using the tools available on Ukiyo-e.org, researchers can now look for a print not just by a title, description, or artist name (there is generally little agreement on the metadata between institutions) and instead find a print by providing just a photo. Additionally, scholars who are researching the manipulation and reuse of the physical woodblocks over time can now more easily locate prints that are derived from the same block but have different imagery. Finally, a tool has been constructed to automatically provide institutions with corrections for their metadata, made possible by finding similar prints and then automatically comparing their associated metadata, looking for differences. All of these tools are able to provide unprecedented improvements to researchers, scholars, and institutions.

The theater photographs work is at a somewhat earlier stage in its development. The project's early experiments with face recognition in the computer vision library OpenCV¹⁰ identified the location of faces within a photograph reliably, but could not be used to suggest whether a face in one photograph belonged to the same person as a face in another. More promising has been the OpenBR¹¹ library from MITRE Corporation which, after "registering" a library of photographs, can quickly return a set of

photographs from the library containing faces that most closely match one depicted in a new photograph. For faces displayed at similar angles and under similar lighting, it performs relatively well, but when the angle changes and additional faces appear in the picture, mistaken identification is more common than success.

Nonetheless, the mistaken identifications are sometimes usefully provocative. False positives reveal similarities among physical characteristics, costumes, and makeup that may not be obvious. For instance, a search using a photograph of young Roddy McDowall as Mordred in the original 1960 Broadway production of Camelot returned (with 100% certainty): Steve Lawrence in a 1967 production of Golden Rainbow, a headshot of actress Sybil White, and a photograph of George C. Scott in Plaza Suite. To the most observers, these faces bear relatively little resemblance, however, to the face recognition algorithm, which looks mostly at the shape and position of the eyes¹², the faces appeared identical. As earlier investigations by Jerome McGann have revealed¹³, this "deformance" of the image by the algorithm may reveal new ways of interpreting the objects. Are there any other similarities (not just visual) among the performers (or the characters they are portraying), that may not have been noticed without the provocation of the algorithm. What do these "fail cases" suggest about the casting practices or makeup design on the mid-20th century Broadway stage?

This paper explores both the promising successes and provocative failures of image analysis tools for humanities research, and suggests future avenues of research the technology makes available to scholars.

References

1. Park, Jung-Ran. *Metadata Quality in Digital Repositories: A Survey of the Current State of the Art*. Cataloging & Classification Quarterly 47, no. 3–4 (2009): 213–228. doi:10.1080/01639370902737240.
2. ukiyo-e.org
3. digitalcollections.nypl.org
4. www.playbillvault.com
5. www.ibdb.com
6. dbpedia.org/About
7. Google, Yahoo, and Microsoft all provide image search engines that are capable of searching millions of images.
8. www.imgseek.net
9. services.tineye.com/MatchEngine
10. opencv.org
11. openbiometrics.org
12. J. Klontz, B. Klare, S. Klum, A. Jain, M. Burge. *Open Source Biometric Recognition*, Proceedings of the IEEE Conference on Biometrics: Theory, Applications and Systems (BTAS), 2013.
13. Jerome McGann and Lisa Samuels *Deformance and Interpretation* www2.iath.virginia.edu/jjm2/old/deform.html (augmented version also available in *New Literary History* 30 (winter, 1999), 25–56

Introducing digital humanities through the analysis of cultural productions

Reyes-Garcia, Everardo

everardo.reyes-garcia@univ-paris13.fr

University of Paris 13

1. Introduction

In this contribution we present a pedagogical approach with the intention to introduce digital humanities to undergraduate students. Our approach may be regarded from three angles: first, as the construction of tailored toolkits of digital methods for

students; second, as a contribution to the analysis of material properties of cultural productions, and; third, as a mid-term strategy to orient students toward design-based learning techniques.

The way in which our pedagogical practice connects the three perspectives is as follows. We consider the realm of cultural productions populated by music albums, films, comic books, TV series, video games, digital art, architecture, industrial design, etc. Now, with the emergence of various kinds of tools and scripts for analyzing media data (text, images, audio, etc.), we select and assemble several of them in a tailored toolkit for studying cultural productions. Then we use the toolkit as a teaching methodology in the classroom. In the mid/long-term, our intention is to move students from the use of tools (as it happens in undergraduate courses) to the creation and design of tools, services and processes (as it happens in postgraduate courses). In the present work, we discuss some experiences with undergraduate students in information and communication sciences at the University of Paris 13.

2. Analyzing cultural productions

Before introducing digital tools and methods in the classroom, we discuss about cultural productions: what they are and why/ how to study them . Within the context of the undergraduate curricula, our course explores the possibilities of digital and connected technologies. Our course tries to complement other types of communicational analysis such as discourse analysis, semiotic analysis, and quantitative methods. From this perspective, our methods put special attention on the analysis of material properties of cultural productions.

In that respect, the analysis of cultural productions deals with tasks such as gathering, documenting, representing, and exploring valuable data about forms, materials, contexts, techniques, themes, and producers of these productions. Cultural objects, or cultural productions, represent the tangible or perceivable result of cultural labor.

In our course, we tackle the analysis of material properties of cultural productions from three dimensions: texts, images, and networks. For practical goals, we first ask students to select a production of their choice: a CD album, a film, a comic book, a series of comic covers, a video game, a music video clip, etc. Then two main types of data are collected. On the one hand, media-based data (texts, images, videos, audios, etc.) and, on the other hand, data about data (metadata) such as years, places, actors, roles, etc. Our toolkit of digital methods is tailored to suit the analysis of each dimension.

Once data has been collected, the next step is to perform information processing techniques in order to generate 'analytical maps', which are the formal outcome of the analysis of material properties of cultural productions. These maps are helpful in the processes of identification of relationships, observation, comparison, evaluation, formulation of hypothesis, verification of intuitions, elaboration of conclusions, and other social sciences methods.

By learning to manipulate tools and studying material features of cultural productions, students generate their analytical maps and use them as a support to reflect on second-order questions: Why the production was made in such a way? Who created it, how, and by which means? Which actors contributed to it and which roles they played? In which manners those actors influenced the final product? How does the production reflect on societal, scientific, temporal, artistic, and geographic aspects of its time?

3. Our approach

3.1. First step: gather data

As we mentioned above, we first ask students to select a cultural production. The selection is free and subjective, it is an individual decision in order to create a comfortable ambient for research. Students are naturally attracted to an artist or film or CD and this might stimulate to dig deeper in the gathering

of data. From another perspective, the choice also reflects ideological presumptions, intuitions and trends in a generation.

For media data, the sources vary according to the choice of the cultural production. In the case of a CD album, for example, texts can be found in the lyrics of songs; for a film it could be the script or even a SRT subtitle file. For comics, it could be the dialog balloons and other paratexts. For images the case is not very different. Images are considered as any graphical information that pertains to the cultural production. CD albums have covers, booklets, etc. Films have frames, posters, etc. Comics have covers, pages, frames, etc.

For data about data (metadata), students use extensively search engines, Wikipedia, specialized online databases (AllMusic , IMDB , etc.) and Google services (Ngram Viewer , Zeitgeist , etc.) to gather data associated with the production: persons and roles (producers, directors, artists, designers, engineers, etc.); years, places, company, label, duration, technical details, etc.

In any case, students take their own decisions about what kind and how large the corpus of analysis should be. This is the reason why we let students to select freely the cultural production, if they like it they can go deeper and construct bigger corpora.

3.2. Second step: analytical map of digital texts

The first type of analytical maps we generate have text as media data input. We mainly rely on four techniques: 1) generating a word cloud; 2) generating a list of word frequencies; 3) generating a word trend graph and identifying the word in context; 4) generating an exploratory visualization of text: a phrase network or an experimental representation of text.

These techniques are coupled with technological tools. We use easy-to-use web-based software. Word clouds are generated via Wordle . A list of word frequencies, a graph of trends, and words in context can be obtained with voyeurtools.org. Finally, exploratory representations of text can be achieved with ManyEyes or other Voyeur tools .

3.3. Third step: analytical map of digital images

The second type of analytical maps we generate have images as media data input. We consider five techniques: 1) extracting the color scheme and listing color values; 2) evidencing shapes; 3) distributing colors according to the RGB color model; 4) generating orthogonal views of video sequences.

As it happens with text, image techniques correspond to specific tools. For technique no. 1 we use the add-on tool Rainbow 1.5.1 available for Firefox . For technique no. 2 we use the online editor Pixlr , specially the filter 'detect contours' combined with an adjustment of brightness and contrast. For technique no. 3 we use the Firefox add-on Color Inspector 3D . For technique no. 4 we use the tool slitscanner.js (only available for HTML5 videos).

3.4. Fourth step: analytical map of digital metadata

The third type of analytical maps we generate have metadata as input. Networks are about rendering evident the relationships between data (for instance, persons involved at some point or playing a particular role in the production) of the cultural object. We work on two techniques: 1) cleaning and preparing data in a spreadsheet; 2) generating and navigating network diagrams.

The first technique is accomplished with Google Spreadsheets, and the second one with ManyEyes.

3.5. Fifth step: analysis of analytical maps to elaborate conclusions

The last part of our approach involves all the analytical maps together. The main goal is to use social sciences methodologies

(observation, comparison, etc.) to elaborate conclusions about the second-order questions: Why the production was made in such a way? How does the production reflect on societal, scientific, temporal, artistic, and geographic aspects of its time?

This last part is most of the time conducted by students themselves or in teams. They often recreate some of the latter steps or they start searching for more resources. In the end, they are free to design a display support for the analytical maps and the conclusions. I, as teacher, do not make suggestions at this stage because students now use more naturally the web as a service.

4. Conclusions and perspectives

We have used our toolkit as teaching strategy for two years and we have documentation on more than 100 student projects. We have collected informal data about student experiences: technical issues, methodology and even cultural trends (for example, most analyzed groups and films). Among our ideas for evaluation and improvement, we foresee: to design higher level courses based on the learning outcomes of this course; to make available a reference manual of DH techniques for students; and, to collaborate closer with other colleagues to complement other types of analysis.

Our toolkit of digital methods is inspired by techniques that come from the domain of text analysis, visual semiotics, and network analysis. Within a digital context, we believe they foster a more scientific web culture as the web is regarded as a platform and service for research. In that manner, the role of the teacher is more to assist students in every step of the analysis and to help identify valuable insights that could only be appreciated through digital methods.

References

- Software Studies Initiative** (2008). *Cultural analytics*. lab.softwarestudies.com/p/cultural-analytics.html (accessed 7 March 2014).
- AllMusic. www.allmusic.com/ (accessed 7 March 2014).
- Internet Movie Database. www.imdb.com/ (accessed 7 March 2014).
- Google Ngram Viewer. books.google.com/ngrams (accessed 7 March 2014).
- Goolge Zeitgeist. www.google.com/zeitgeist/ (accessed 7 March 2014).
- Wordle. www.wordle.net/ (accessed 7 March 2014).
- >ManyEyes. www-958.ibm.com/software/analytics/manyeyes/ (accessed 7 March 2014).
- Voyeur Tools. Online: hermeneuti.ca/voyeur/tools (accessed 7 March 2014).
- Rainbow 1.5.1. addons.mozilla.org/en-US/firefox/addon/rainbow-color-tools/ (accessed 7 March 2014).
- Pixlr. pixlr.com/editor/ (accessed 7 March 2014).
- Color Inspector 3D. addons.mozilla.org/En-us/firefox/addon/color-inspector-3d/ (accessed 7 March 2014).
- Hwang, S.** (2013). *Slitscanner.js*. sketches.postarchitectural.com/slitscanner/ (accessed 7 March 2014).
- Reyes, E.** (2013). *Culture numérique*. bit.ly/1fMD0gU (accessed 7 March 2014).

The Story of Stopwords: Topic Modeling an Ekphrastic Tradition

Rhody, Lisa

lmrhody@gmail.com

Roy Rosenzweig Center for History and New Media

Introduction

This paper will argue that removing high-frequency, low-semantic weight words from topic models of poetry corpora improves the coherence of Latent Dirichlet Allocation (LDA) topics and addresses reasonable concerns by some literary scholars that removing such language undercuts the methodology's value as a mode of literary inquiry. Exposing technical and theoretical decisions made while topic modeling 4,500 English language poems, this paper demonstrates how words such as "look" and "saw" remain influential and semantically present in document to topic distributions. Finally, it suggests that literary scholars will need a different hermeneutic approach to topic models of poetic corpora that better accounts for ambiguity and figures of speech.

Background

Ekphrasis—poems to, for, and about the visual arts—offers a wealth of opportunities to ask familiar humanities questions about canon-formation, literary tradition, and genre definition, and at the same time affords avenues for the advancement or refinement of methods and tools in the field of digital humanities. The story of the ekphrastic tradition, and women's relationship to that tradition, is in many ways the story of data collection and curation. In his influential essay on the genre, W. J. T. Mitchell radically shifts critical studies of poetic engagements with images away from metaphorical comparisons by arguing that ekphrasis activates historical and ideological oppositions between the linguistic and spatial arts as a staging of anxieties about "otherness."¹ Mitchell goes on to explain that the "treatment of the ekphrastic image as female other is commonplace in the genre" (164).

To date, Mitchell's theorization of ekphrasis as social contest remains a powerful influence on our critical approaches to how the genre operates because it pushed beyond previous studies that simply compared the two arts formally. Mitchell's "cannon," however, consists of four poems, all by male poets: Wallace Stevens's "Anecdote of a Jar;" William Carols Williams's "Portrait of a Lady," John Keats's "Ode on a Grecian Urn;" and Percy Bysshe Shelley's "On the Medusa of Leonardo Da Vinci in the Florentine Gallery." Though recent scholars--such as Elizabeth Loizeaux², Jane Hedley³, and Barbara Fischer⁴—have challenged the limitations of Mitchell's model, two challenges have stymied its revision: identifying genre conventions as succinct and intuitive and surveying a much larger collection of ekphrastic examples.

Literary scholars respond to questions of genre by creating models. For example, when Mitchell describes the "suturing of gender stereotypes" onto the "interworkings of ekphrasis," he does so by creating a model he calls the ekphrastic triangle. Mitchell's triangle stages a relational exchange between a poet-speaker, a feminized art object and the reader, in which the speaker instructs the reader to "look" and "see," cautions him against silence and stillness, and confides a desire to ravish the feminized image. This presentation demonstrates that ekphrasis's re-deployment of multiple discourses expands our model of ekphrasis as a network by situating individual poems among multiple, ongoing, and constantly changing discourses within the topic model.

Opportunity

Computational tools, such as topic modeling, have the potential to help literary scholars redefine the limits of reading distance, not because they read better than humans but because computers compute better than humans. Soon after he first coined the phrase "distant reading," Franco Moretti claimed that distance is a "condition of knowledge"—one limit of many.⁵ ⁶ Leveraging the strengths of technology to broaden the reach, scale, and scope of our exposure to ekphrastic poetry improves our capacity to view the ekphrastic tradition on a much larger scale. Although the history of ekphrasis—as the hostile, gendered contest between speaking male subjects and silent female objects—seems particularly inhospitable territory for women poets, many acclaimed poems by women in the 20th century participate in the genre, including

Elizabeth Bishop's "Poem," Anne Sexton's "Starry Night," Jorie Graham's "San Supolcro" and Elizabeth Alexander's "The Venus Hottentot." Topic modeling's generative, probabilistic, and non-semantic methods offer a promising opportunity to revise our critical understanding of ekphrasis. Removing words such as "see," "look," and "say" from ekphrastic poems offers opportunities to refocus our critical lens on other language patterns that have been overlooked by human pattern recognition due to the high frequency of "look" and "see" throughout the ekphrastic canon.

Methods

This paper asks: if we can discern salient questions about the ekphrastic tradition that computer reading is designed to address, how might we respond to Adrienne Rich's call for "re-vision," whereby learning to see what we already know differently is an act of survival? Since we know that computers are not the same kind of readers that humans are, it is important to be aware of the "conditions" that shape the ways computers and humans read. Questions that require attention to quantity, scope, and scale are particularly suited to computation, and computation helps adjust the aperture on the lens that a scholar can have of a corpus of texts. The challenge, however, is to continue to refine our understanding of what questions might be most fruitfully asked with an awareness of computational and human conditions.

In the study of ekphrasis, for example, interpretive stress is placed on high-frequency words that, in prose, hold relatively little semantic weight--particularly words such as "look" and "see." Distinct for its highly concentrated language, poetry places an increased degree of significance on even the poet's "smallest" word choices. Preprocessing texts for topic modeling in Mallet strips documents of upper and lower case letters, removes line breaks and enjambments, deletes high-frequency words, including articles, prepositions, pronouns, conjunctions, and common verbs--like "is," "are," and "were"--and turns documents into strings of sequential words that no longer bear the same syntactical relationships they once did. Given this, how can a methodology that requires radical decomposition of a poem's linguistic meaning offer valuable insights into exploring texts? For example, heavy with subtext, the first line of Robert Browning's "My Last Dutchess" would be removed in its entirety from the text of the poem through preprocessing using the default stop list available in Mallet.⁷ If the computer doesn't value such words, could the model it produces still be useful to someone interested in ekphrasis?

To determine whether or not LDA could still produce models that would be useful to the study and revision of the ekphrastic tradition, four experiments were performed on a dataset of 4,500 English language poems using four different preprocessing techniques: 1.) removing no words before creating the model 2.) removing only 50% of the Mallet stopwords 3.) removing all the Mallet stopwords except a small group that relate to ekphrastic conventions (eg. look, see, saw, seen) and 4.) removing all the words on the suggested Mallet stoplist. After preprocessing, each dataset was treated identically, producing 40 topic models. This paper addresses the decision process for assembling each list and points to where the lists can be found online.

Discoveries

Contrary to expectation, the model with the greatest topic key word distribution coherence was the model with the most stopwords removed. Although ekphrastic poems beseech their readers to "look" and to "see" more clearly, the ekphrastic poems themselves surface more coherently in models where the words "see," "look," and "still" are removed. Like ghosts in the model, similar topics focused on looking and describing form even when specific words referring to the activity are no longer present, and the model's prediction of topical distribution more accurately reflects human estimation of the numbers of "ekphrastic" versus "non-ekphrastic" poems were included in the dataset to begin with. In fact, topics in the lightly edited

stoplist test are reflected in the model where all stopwords are removed, but the list of most likely words associated with each topic is less coherent in the former than the latter's keyword list.

Using specific examples of topics created with models using various stopword lists, this paper tells the story of ekphrasis as it is told through stopword filters, as topical coherence rises, the questions one might ask about the corpus changes. Concentrating on terms such as "still," "look," and "see," this paper will demonstrate how LDA identifies topics where issues of vision, description, and color become refined as the words that directly refer to the act of observation are removed. Visualizations of relationships between poetic topics and topic word and document distributions will also reveal the ways in which latent patterns of words that have been removed from the corpus can still be evident in topic formation.

Copies of the model keyword distribution lists and document to topic distributions are available by request.

Conclusion

This paper uncovers the complementary theoretical and methodological decisions required in order to approach questions of tradition and canon formation with topic modeling corpora of poetry. An act of critical "deformance," topic modeling uncovers differences between Keats' "still unravish'd bride of quietness" and Carol Snow's "Positions of the Body" even without many of the words that scholars have argued were critical identifying features of the genre. Such discoveries prompt us to return to close readings of ekphrastic poems with new questions about the conventions of a genre evident in Western letters since Homer's first description of Achilles' shield in the *Illiad*.

References

1. Mitchell, W.J.T. (1994). *Ekphrasis and the Other*. Picture Theory. Chicago: University of Chicago Press. Print.
2. Loizeaux, Elizabeth. *Twentieth Century Poetry and the Visual Arts*. (2008). Cambridge, UK: Cambridge UP. Print.
3. Hedley, Jane, Nick Halpern, and Williard Spiegelman. (2009). *In the Frame: Women's Ekphrastic Poetry from Marianne Moore to Susan Wheeler*. Newark, DE: University of Delaware Press. Print.
4. Fischer, Barbara K. (2006). *Museum Meditations: Reframing Ekphrasis in Contemporary American Poetry*. 1st ed. New York: Routledge. Print.
5. Moretti, F. (2000). *Conjectures on World Literature*. New Left Review 1. newleftreview.org/II/1/franco-moretti-conjectures-on-world-literature
6. Moretti, F. (2005). *Graphs, Maps, Trees: Abstract Models for a Literary History*. Verso.
7. McCallum, Andrew Kachites. (2002). *Mallet: A Machine Learning for Language Toolkit*. Web. 31 May 2013.

Play as Process and Product: On Making Serendip-o-matic

Ridge, Mia

Mia.Ridge@open.ac.uk
The Open University, United Kingdom

Croxall, Brian

brian.croxall@emory.edu
Emory University, United States of America

Papaelias, Amy

papaelia@newpaltz.edu
State University of New York at New Paltz, United States of America

Kleinman, Scott

scott.kleinman@csun.edu
California State University, Northridge

Abstract

"Play" is not a foreign concept to the digital humanities nor to the Digital Humanities Conference. Indeed, there have been a number of presentations over the years that focus on the ludic interactions between player and game in virtual worlds; that consider game design; that attempt to preserve experiences of gameplay; that view play as integral for the creation of new texts or interpretation of already existing ones; or that philosophically connect play to pedagogy. There have even been panels that examine the literal mechanics of play in the context of video. But while play has been central to these discussions, it has most often been treated *very seriously*.

In contrast, this presentation will highlight a project that placed playfulness at the center of its development and of its final product. Serendip-o-matic (serendipomatic.org) is a "serendipity engine" that connects large texts (e.g., entire articles or syllabi) or personal research libraries to digital materials located in libraries, archives and museums around the world. It recreates the surprising discoveries that frequently accompany research. Serendip-o-matic was built in less than five days as part of One Week | One Tool (OWOT, oneweekonetool.org). OWOT is a playful departure from traditional institutes funded by the U.S.'s National Endowment for the Humanities (NEH) in that the outcomes are largely determined by its participants rather than by the institute's organizers. And the team that designed and built Serendip-o-matic decided to place play at the center of its process and final product.

"Play" has been defined as the "free movement within a more rigid structure". In the context of tool building, a playful process means dodging the rigid structures that so often define the tool development process. A digital humanities "barn-raising" destabilizes the established ways in which digital humanities tool building projects are traditionally formulated, facilitating a playful process. The impending deadline—just. one. week! —and the rapid-fire completion of tasks and decision-making made the building experience intense and challenging. This presentation will discuss the various ways in which the participants chose to frame the experience, from hack days to games to reality TV. Although it was a very real possibility that Serendip-o-matic might fail to launch on time (if at all), the artificial construction of the challenge—as an NEH institute, wherein the main goal was for participants to learn something—made the consequences of such failure less severe. Separated from the responsibilities of daily life, the immersive format of OWOT offered the team the opportunity to take more risks and engage in a playful practice of ideation, conceptualization, and making.

Play can also be a significant method for creating engaging user-experiences in design. This technique has been employed in a variety of user-experience environments, particularly those aimed at engaging the public with humanities and cultural heritage materials, such as the Tate Museum's "Magic Tate Ball" and the Wellcome Collection's "High Tea". The playfulness of Serendip-o-matic is easily visible upon visiting the project's website. The logo resembles a Rube Goldberg machine, and the colors and iconography throughout the site were chosen and created specifically to evoke the lively nature of the tool itself. The text throughout the site also broadcasts a playful attitude. For example, users are asked not to "select" or "upload" a "document" but instead to simply "grab some text." Much of the work of the Outreach team during the week was spent crafting this text as well as the name for the tool, with the goal of preserving the playful feel of the entire enterprise.

In this presentation, members of the OWOT team will report on creating a playful digital humanities tool. In addition to discussing the rapid and ludic development atmosphere and the choices made to shape user experience, the presentation will discuss the playfulness that was encoded in the tool's discovery algorithm, which uses named-entity recognition and text parsing to create a 'magic machine' that queries various collection APIs. We will also outline steps that were taken to create a playful voice throughout the deployment of the tool, including the creation of a mascot for the tool (the "Serendhippo") and

rules for public engagement for the mascot as well as the whole team. We will also report on the challenges of a playful project, one of which includes the consideration of how such playful work should be evaluated in the academy.

In conclusion, we will argue in favor of the benefits for incorporating more "playful work" in the context of academic research and scholarship and the significant role played by the interface and user experience design skills of several team members. As current digital humanities work relies on collaborative environments (including hackathons, maker spaces, maker challenges, etc.), opportunities like One Week | One Tool provide a space for playful work to encourage more creative risk-taking and engaging user-experiences within the context of digital humanities scholarship and practice.

References

- Kirschenbaum, M., et al.** (2009). "Twisty Little Passages Not So Much Alike: Applying the FRBR Model to a Classic Computer Game." Digital Humanities 2009. University of Maryland, College Park. 23 June 2009.
- Jones, S. E.** (2009). "The Social Text as Digital Gamespace: or, what I learned from playing Spore." Digital Humanities 2009. University of Maryland, College Park. 25 June 2009.
- Bonsignore, B., et al.** (2011). *The Arcane Gallery of Gadgetry: A Design Case Study of an Alternate Reality Game*. Digital Humanities 2011. Stanford University. 21 June 2011. Poster presentation.
- Lowood, H. and J. McDonough.** (2009). "The Open Archival Information System Reference Model vs. the BFG 9000: Issues of Context and Representation in Game Software Preservation." Digital Humanities 2009. University of Maryland, College Park. 23 June 2009.
- Kraus, K., R. Donahue, and M. Winget.** (2009). "Game Change: The Role of Professional and Amateur Cultures in Preserving Virtual Worlds." Digital Humanities 2009. University of Maryland, College Park. 23 June 2009.
- McClure, D. W.** (2012). "Exquisite Haiku: Experiments with Real-Time, Collaborative Poetry Composition." Digital Humanities 2013. University of Nebraska-Lincoln. 17 July 2013. <http://dh2013.unl.edu/abstracts/ab-155.html> (accessed 29 October 2013).
- Drucker, J. and J. McGann** (2002). "Ivanhoe: A Game of Interpretation." Digital Humanities 2002. University of Tübingen. July 2002. 67.207.129.15:8080/dh-abstracts/view?docId=2002_panel_060_rockwell.xml;query=play;brand=dh-abstracts (accessed 29 October 2013).
- Rockwell, G.** (2002) "Is Gaming Serious Research in the Humanities?" Digital Humanities 2002. University of Tübingen. July 2002. 67.207.129.15:8080/dh-abstracts/view?docId=2002_panel_060_rockwell.xml;query=play;brand=dh-abstracts (accessed 29 October 2013).
- Harris, K. D.** (2011). "Pedagogy & Play: Revising Learning through Digital Humanities." Digital Humanities 2011. Stanford University. 20 June 2011. <http://dh2011abstracts.stanford.edu/xml/view?docId=tei/ab-242.xml;query=harris%20pedagogy;brand=default> (accessed 29 October 2013).
- Croxall, B.** (2012). "Courting 'The World's Wife': Original Digital Humanities Research in the Undergraduate Classroom." Digital Humanities 2012. University of Hamburg. 18 July 2012.
- McDonald, J. L., A. K. Melby, and H. Hendricks.** (2010). "Standards, Specifications, and Paradigms for Customized Video Playback." Digital Humanities 2010. King's College London. 10 July 2010.
- Salen, K. and E. Zimmerman.** (2003). *Rules of Play: Game Design Fundamentals*. Cambridge: The MIT Press.
- Korhonen H., M. Montola, and J. Arrasvuori.** (2009). "Understanding Playful Experiences Through Digital Games." In proc. of the 4th International Conference on Designing Pleasurable Products and Interfaces, DPPI 2009. research.nokia.com/files/p274%20-%20Korhonen.pdf (accessed 29 October 2013).
- The Tate Museum.** (2012). "Magic Tate Ball app." Blogs & Channels. <http://www.tate.org.uk/context-comment/apps/magic-tate-ball> (accessed 1 November 2013).

Burgoyne P. (2012). "The Magic Tate Ball." CreativeReview Blog. 15 May 2012. www.creativereview.co.uk/cr-blog/2012/may/magic-tate-ball (accessed 30 October 2013).

Wellcome Collection. ([2010]). *High Tea*.

hightea.wellcomeapps.com (accessed 1 November 2013).

Birchall, D. and M. Henson. (2011). "High Tea Evaluation Report." museumgames.pbworks.com/w/file/fetch/44614076/HighTeaEvaluationReport.pdf (accessed November 1 2013).

Goldberg, R. et al. (2013). *The Art of Rube Goldberg: (A) Inventive (B) Cartoon (C) Genius*. New York: Abrams ComicArts.

Harddrive Philology: Analysing the Writing Process on Thomas Kling's Archived Laptops

Ries, Thorsten

Ghent University, Belgium

The proposed talk will discuss the application of forensic computer science tools and methods to born digital documents and parts of archives, focusing on the philological benefit for genetic scholarly editions and the *critique génétique* on the one hand as well as on issues of sane archiving and representation of the digital record in a scholarly edition on the other. In the course of the talk, the conceptual impact of this digital forensic perspective on the term 'document', on our concepts of the 'materiality of the historical record' and on 'textual genetics' will also be discussed.

Exemplary subject matter of the inquiry will be a selection of recovered materials from harddrives in the Thomas Kling archive which represents the range of retrievable transitory 'genetic' textual material, e.g. recoverable documents, temporary files, memory fragments and several disk operation artifacts on multiple generations of operating systems reaching back from Microsoft Windows® XP to 3.11. The harddrives have been forensically imaged for longterm preservation by the author of this paper and by a forensic laboratory recently, and it will be the first time these findings are being publicly discussed.

Thomas Kling (1957-2005) was one of the most influential contemporary poets of the last 30 years in the German-speaking countries. The historical, documentary quoting technique as well as his blending of poetry and performance inspired numerous other authors of his generation. Furthermore, his poetry is enormously aware of the effect that historical media development has on perception, on the way how storage media influence the concept of history and on poetic language itself (s.a. Trilcke 2012). As early as 1985, he wrote: "[...] and everyone knows: from now on, we cast out poetry on floppy disks only, sure thing." (Thomas Kling: *Die verplemperten Sprachen*; Wehr et al. 2012: 13) At the same time, his work reflects not only the medial blending of historical levels and documents as „sondage“, but also the threat of losing the „burning archive“: „It is the tongue-loss. Everything is archive, everything is about to become archive and end up in smoke.“ (Kling 2001: 111) Fortunately, this does not apply to his own archive, nor to the harddrives of his last three laptops, all of which are being kept in the collections on the Raketenstation Hombroich (Scharfschwert 2012).

After a quick introduction to general archival aspects and methods of forensic work with bitstream-preserving images and the several levels of analysis (different kinds of file recovery, drive slack analysis, save operation artifacts, restore points etc., s.a. Ries 2010, Kirschenbaum et al. 2010, Reside 2011), the talk will discuss a couple of example findings from the Thomas Kling harddrive platters to show in which – sometimes surprising – places of these 'real life' case systems textual variants and fragments of poetic draft material actually reside. Possible candidates for this part are digital fragments of the *dossier génétique* to the poems *third cartography* and its *abdomen in constant movement* (selection to be finalised). The example materials will show Thomas Kling as one of the German

poets who relatively early embraced IBM-compatible personal computers with Microsoft Windows®, Word® as a writing tool and used it for most of the draft process, going back and forth between the digital document and corrected printouts after a conception phase on notebook and manuscript pages.

Discussion will show to what extent the philological interpretation of these findings depend on the specifics of the operating system- and application context and that we sometimes have to deal with 'artificial' evidence. Furthermore, the range of variation in terms of completeness of the textual record will be mapped. A tentative genetic close reading of the fragments will also show how the reader's view on the writing process necessarily shifts, coming from a manuscript perspective, as one reads e.g. recovered digital born memory snapshot items.

Looking at the materials from a scholarly editing point of view, questions arise how these should be represented in a genetic edition, e.g. which metadata has to be included; in which meaningful way can the commentary cover the technical context of the mechanisms of historic software, and how is redundancy of the digital record to be dealt with? (s.a. Pierazzo 2011) How is the 'materiality' of the materials to be represented? (s.a. Ries 2010, Kirschenbaum 2006)

In this sense of the meaning, this talk aims to help „empowering“ scholarly editors, philologists and scholars as readers of future genetic editions to deal with digital heritage collections and digital born documents and material as part of scholarly editions.

Delivering this talk in German would be an obvious choice, as the discussed archive materials by Thomas Kling are also in German and I am a native speaker. However, the talk can be held in English as well to reach out for the international audience. Regarding the advantages of both options, I would like to leave the choice of language for this talk to the conference board.

References

- Gitelman, Lisa** (2006). *Always Already New. Media, History, and the Data of Culture*. Cambridge, London: MIT Press; 2006.
- Kirschenbaum, Matthew G., Richard Ovenden, Gabriela Redwine** (Eds.) (2010): *Digital Forensics and Born-Digital Content in Cultural Heritage Collections*. Council on Library and Information Resources Washington, D.C. December 2010 (CLIR publication, no. 149).
- Kling, Thomas** (2001): *Rhapsoden am Sepik*. Botenstoffe. Cologne 2001.
- Pierazzo, Elena** (2011). *A rationale of digital documentary editions*. LL&C 26.4 (2011), 463-77.
- Reside, Doug** (2011): "No Day But Today": A look at Jonathan Larson's Word Files. NYPL Blog, April 22, 2011, URL: www.nypl.org/blog/2011/04/22/no-day-today-look-jonathan-larsons-word-files, accessed: October 30, 2013.
- Ries, Thorsten** (2010). "die geräte klüger als ihre besitzer" *Philologische Durchblicke hinter die Schreibszene des Graphical User Interface. Überlegungen zur digitalen Quellenphilologie, mit einer textgenetischen Studie zu Michael Speiers "ausfahrt st. nazaire"*. editio 24.1 (2010), 149-199.
- Ries, Thorsten** (2012). Review: **Willard McCarty** (Ed.). *Text and Genre in Reconstruction* (2010)
- Scharfschwert, Alena** (2012): *Das eingepflegte Archiv. Bericht über die Erschließung des Thomas Kling-Nachlasses*. Frieder von Ammon, Peer Trilcke, Alena Scharfschwert (Eds.) Das Gellen der Tinte. Zum Werk Thomas Klings. Göttingen: V&R unipress 2012, 383-400.
- Stüssel, Kerstin, Gabriele Wix, et al.** (Eds) (2013): *Zur Leitcodierung. [Manhattan Schreibszene; Katalog zur Ausstellung "Thomas Kling: geschmolzener und/ wieder aufgeschmolzener text" 2013/14 im Universitätsmuseum der Universität Bonn]* Göttingen: Wallstein 2013.
- Trilcke, Peer** (2012): *Historisches Rauschen*. Das geschichtslyrische Werk Thomas Klings. [The Historical Poetry of Thomas Kling]. PhD thesis Göttingen University 2011/12, URL: hdl.handle.net/11858/00-1735-0000-0006-AEDE-3, accessed: October 30, 2013.

Norbert Wehr, Ute Langanky, Marcel Beyer (Eds.) Thomas Kling (2012). *Das brennende Archiv. Unveröffentlichte Gedichte, Briefe, Handschriften und Photos aus dem Nachlass [...].* Berlin: Suhrkamp 2012.

A Network Analysis Approach of the Venetian Incanto System

Rochat, Yannick
yannick.rochat@epfl.ch
EPFL

Fournier, Melanie
melanie.fournier@epfl.ch
EPFL

Mazzei, Andrea
andrea.mazzei@epfl.ch
EPFL

Kaplan, Frédéric
frédéric.kaplan@epfl.ch
EPFL

The Venetian maritime empire is the subject of numerous works and monographs (e.g. Ercole 2006¹, Lane 1973², Luzzatto 1941³). This paper focuses on the period between the end of the 13th century and the fall of Constantinople in 1453. During that period the Venetian state set up seven regular shipping lanes, linking the Republic of Venice with the oriental and the occidental Mediterranean basins, the Black Sea, England and Flanders. Special warships—called galleys—were readapted to perform commercial duties during peacetime on these shipping lanes. Every year, the Venetian Republic organized an auction system—the Incanto—to assign the commercial space on these ships. Subsequently the Senate was in charge of determining the mandatory stopovers, duration of the call, date of departure and date of return to Venice. All of this precise information was recorded in the Venetian official administrative documents.

Several authors have tried to reconstruct the Incanto system from the highly detailed information contained in these administrative documents. In 1961, Tenenti and Vivanti produced a series of chronological maps showing the evolution of the lanes year by year. Unfortunately, their model of the archives is not available for further investigation. More recently, Doris Stöckly extracted from the Venetian state archives—and other sources—a detailed list of all the information related to the ships on a year by year basis. She published her analysis in a monography (Stöckly 1995⁴). The compiled tables appear as appendices to her Ph.D thesis; and are only available in printed form (see figure 1).

For this work, we take these printed tables, digitize, automatically transcribe and structure them. We perform new analyses of the structure and evolution of the Incanto system. Our ambition is to go beyond the textual narrative or even cartographic representation to perform a network analysis which potentially offers a new perspective on this maritime system.

Method

Step 1 : From Printed Tables to Structured Data

The first step of our project was the transformation of the appendices into structured data ready for analysis. We scanned these documents and processed each page using a specifically designed pre-processing pipeline, aimed at improving the quality and highlighting the structure of the scanned images. The pre-processing step included several computer vision-based procedures, serving two main purposes: the adjustment of moderate rotations introduced by the scanning process and the removal of noisy components that may disturb the recognition process. To explicit the structure of the table, we

elaborated a method based on horizontal and vertical projection profile that automatically fit rows and columns of the document table. This grid was then used in conjunction with Optical Character Recognition Software (ABBYY Fine reader). We extracted 1480 lines of data. Each line matches a galley and includes the following information: name of the line, year, number of ships, stopovers, and optionally duration of stay.

1340	d	gal	p	FC - Mo	Dardanelle + E	wsd	sep7	55	wsd4	Closure/City/City/Jerusalem/15/May/grip/0ip	M2258_943/0ip
1340	ty	gal	p	N/1/Ga	Babylonia	wsd	sep7	87 fe	wsd4	Closure/City/2/City/May/grip/0ip	M2258_947/0ip
1340	ty	gal	p	N/1/P	London	wsd	sep7	90	wsd4	Closure/City/2/City/May/grip/0ip	M2258_947/0ip
1345	ty	gal	p	Z + M	Michel	wsd	sep7	91	wsd4	Closure/City/2/City/May/grip/0ip	M2258_947/0ip
1346	ty	gal	p	N/2/Ga	Dardanelle	wsd	sep7	92	wsd4	Closure/City/2/City/May/grip/0ip	M2258_947/0ip
1346	ty	gal	p	N/2/M	Reude	wsd	sep7	94 fe	wsd4	Closure/City/2/City/May/grip/0ip	M2258_947/0ip
1348	ty	gal	p	M/1/Mc	Zembla	wsd	sep7	94	wsd4	Closure/City/2/City/May/grip/0ip	M2258_947/0ip
1348	ty	gal	p	Luchas	ZemblaZero	wsd	sep7	95	wsd4	Closure/City/2/City/May/grip/0ip	M2258_947/0ip
1349	ty	gal	p	Barbaria	Sarawat	wsd	sep7	96	wsd4	Closure/City/2/City/May/grip/0ip	M2258_947/0ip
1349	ty	gal	p	M/1/P	Corsica/S	wsd	sep7	95	wsd4	Closure/City/2/City/May/grip/0ip	M2258_947/0ip
1347	ty	gal	p	Th/1/M	Babylonia	Gustavus M	jan20	155 fe	sep10	Caled/City/City/Celte	M2425/212/0ip
1347	ty	gal	p	P/1/M	Contras	Gustavus M	jan20	156	sep10	Caled/City/City/Celte	M2425/212/0ip
1347	ty	gal	p	M/1/M	Barbary	Gustavus M	jan20	157	sep10	Caled/City/City/Celte	M2425/212/0ip
1347	ty	gal	p	Zeneca	Mars	Gustavus M	jan20	158	sep10	Caled/City/City/Celte	M2425/212/0ip
1347	ty	gal	p	Bordigia	Bordigia	Gustavus M	jan20	159	sep10	Caled/City/City/Celte	M2425/212/0ip
1347	ty	gal	p	M/1/M	Mezo	Gustavus M	jan20	160	sep10	Caled/City/City/Celte	M2425/212/0ip
1347	ty	gal	p	Mezo	Contras J	Gustavus M	jan20	160 fe	sep10	CP/DMN/May/grip/0ip	M2391/MJ/2024
1347	ty	gal	p	Peru	Contras J	Gustavus M	jan20	160	sep10	CP/DMN/May/grip/0ip	M2391/MJ/2024
1347	ty	gal	p	Dard	Monserrat	Gustavus M	jan20	160	sep10	CP/DMN/May/grip/0ip	M2391/MJ/2024
1347	ty	gal	p	Nicola	Cepolla	Gustavus M	jan20	160	sep10	Assendelft/R/sea/Cele/GD/2104	M2417/20/0ip
1347	ty	gal	p	Asellio	Asellio	Gustavus M	jan20	161	sep10	Assendelft/R/sea/Cele/GD/2104	M2417/20/0ip
1347	ty	gal	p	Barbary	Asellio/Circe	Gustavus M	jan20	162	sep10	Assendelft/R/sea/Cele/GD/2104	M2417/20/0ip
1347	ty	gal	p	Mareto	Cepolla	Gustavus M	jan20	163	sep10	Assendelft/R/sea/Cele/GD/2104	M2417/20/0ip
1347	ty	gal	p	P/1/L	Moorigo	VATAS	fev20	70	sep10	Bar/Akkad/4/Bre	M241/5/0ip/1v
1348	sy	gal	p	T/1/M	Babylonia	Corsica Max	dec0	127.5	fev1/16	Chypriat/0ip	M241/5/0ip/2
1348	sy	gal	p	Re/1/M	Babylonia	Corsica Max	dec0	128	fev1/16	Chypriat/0ip	M241/5/0ip/2
1348	sy	gal	p	Denava	Phanaxo	Corsica Max	dec0	129.5	fev1/16	Chypriat/0ip	M241/5/0ip/2
1348	sy	gal	p	Constantino/Pisan	Corsica Max	Corsica Max	dec0	130.5	fev1/16	Chypriat/0ip	M241/5/0ip/2
1348	sy	gal	p	M/1/M	Tritone/John M	Corsica Max	dec0	130.5	fev1/16	Chypriat/0ip	M241/5/0ip/2
1348	sy	gal	p	Scicli	Venice/Ven	Corsica Max	dec0	130.5	fev1/16	Chypriat/0ip	M241/5/0ip/2
1348	sy	gal	p	M/1/M	Corsica SP	Venice/Ven	dec0	130	fev1/2	Cela/Akkad/Cele/Cels	G251/5/0ip/2
1348	al	gal	p	Nicola	Cepolla	Venice/Ven	dec0	50	fev1/2	Cela/Akkad/Cele/Cels	M241/5/0ip/42
1348	rs	gal	p	Schutte	Lombardo	Monserrat	dec0	110.5	fev1/16	Nigrop/CPN/1/Sea/Cels/4/0ip	M241/5/0ip/42,M2353
1348	rs	gal	p	Jacob	Amredd	Monserrat	dec0	110	fev1/16	Nigrop/CPN/1/Sea/Cels/4/0ip	M241/5/0ip/42
1348	rs	gal	p	St. Iwan	St. Iwan	Monserrat	dec0	110	fev1/16	Nigrop/CPN/1/Sea/Cels/4/0ip	M241/5/0ip/42
1349	norm	gal	p	St. Iwan	Storaco SM	Falena Prod	dec27	92	fev1/16	Ragusa/Adm/Mar/May/grip/0ip	M2071/9/0ip
1349	norm	gal	p	M/1/F	London+HMS	Falena Prod	dec27	92	fev1/16	Ragusa/Adm/Mar/May/grip/0ip	M2071/9/0ip
1349	norm	gal	p	Hercules	Mols + S/S	Falena Prod	dec27	93	fev1/16	Ragusa/Adm/Mar/May/grip/0ip	M2071/9/0ip
1349	norm	gal	p	St. Iwan	St. Iwan	Falena Prod	dec27	93	fev1/16	Ragusa/Adm/Mar/May/grip/0ip	M2071/9/0ip
1349	norm	gal	p	La P/9	Zara	Falena Prod	dec27	95	fev1/16	Ragusa/Adm/Mar/May/grip/0ip	M2071/9/0ip
1349	ty	gal	p	P/1/Ma	Monserrat	Falena Prod	dec27	90	mar15	MogAqj/Cele/Mons/Ragusa IC	M2414/9/0ip/1v

Fig. 1: Excerpt of the extracted data from Doris Stöckly Ph. D thesis appendix.

Step 2 : From Structured Data to Networks

We transformed the resulting table into a network. First, we applied a set of rules in order to clean the data. Then, we removed the stops marked as “ facultative ”. The stops mentioned without any temporal detail were considered as equal to one day—the shortest unit of time. Names of places and geolocations were standardised using a spatial database of Ancient Ports and Harbours based on Harvard's DARMC⁵ and the Pleiades data⁶. We grouped the stopovers under two generic labels for Crete and for Cyprus.

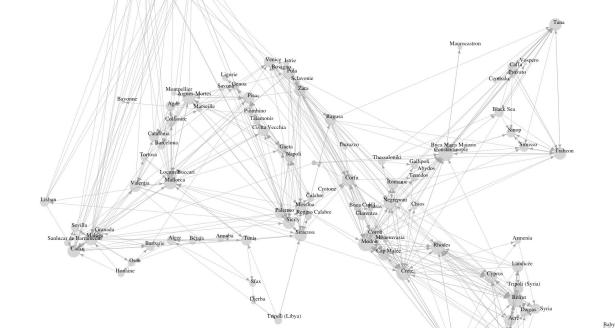


Fig. 2: The 170 years of the Incanto system visualised as a network.

We decomposed—using an R script—the structured table into individual segments made of paired consecutive stopovers. By connecting these directed segments, we created a global directed network encoding 170 years of navigation (see figure 2). The vertices of this network represent all the ports and places mentioned for this period. The size of the nodes is proportional to the sum of in- and out-degree measures of the node. The arcs represent maritime traffic. Two attributes are associated to each arc: one for the year of the trip and another one reporting the number of ships in each convoy.

From the global network, we produced separated subnetworks corresponding to each year of navigation. These

subnetworks inherit their attributes from the main network: the number of ships and days. In figure 3, we illustrate evolution and dynamics of the Venetian maritime routes for the three years before and the three years after the Chioggia war (1351-1354) between Venice and Genoa.

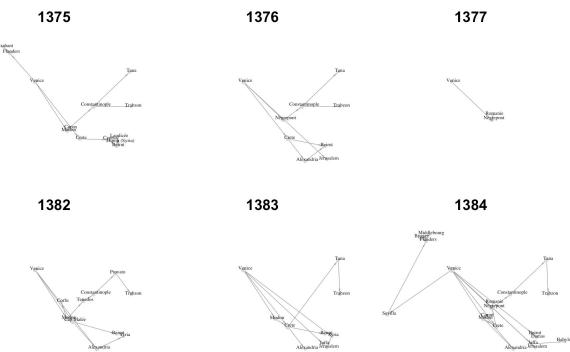


Fig. 3: Network visualization of six years of maritime routes before and after Chioggia war (1377-1381)

Network Analysis: Crete vs. Cyprus

We focused our investigation on two particular islands located in the oriental basin of the Mediterranean Sea: Crete and Cyprus. After its acquisition by the Venetian empire and for 460 years, Crete was a fundamental naval base in terms of localisation, logistics and safety (Dudan 2006, Major 1989). Cyprus had a similar strategic position; it was an intermediary stop and became part of the Venetian empire in 1489.

Based on the network extracted from the *Incanto* dataset, we computed a measure of commercial betweenness of the islands of Crete and Cyprus. In figure 4, we show its time evolution in the period comprised between 1283 and 1453. We highlight three patterns emerging from the computation of this measure and interpret them using three events in the maritime history of Crete and Cyprus.

The first time histogram contains a blue box encapsulating that measure on Crete between 1344 and 1377. During that period, the maritime traffic density increased because of the reopening of the Alexandria lane, as Crete was the last stopover for all the convoys heading to Egypt. It is interesting to compare this change with the increase of commercial betweenness, as highlighted in the figure 4.

In the second time histogram, two red boxes highlight two historical events related to Cyprus maritime traffic. The first one reflects the betweenness of Cyprus as an important stopover on the way to Armenia (1283 - 1338) (Balard 1987). During this period the measure of betweenness naturally skyrockets, as the island had acquired a strategic position as a maritime hub. On the contrary, the second box shows very low measures of betweenness; corresponding to moderate maritime traffic. This was due to the fact that the Senate of Venice reorganised the commercial exchanges by opening a new lane towards Beirut. During this period (1375 - 1444), Cyprus lost its strategic position for maritime activity directed towards Syria and Egypt.

One can notice that the re-opening of Alexandria as destination for Venetian navigation (1344) had the opposite impact on the maritime traffic passing through Cyprus and Crete.

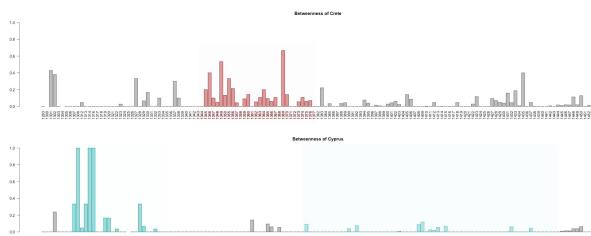


Fig. 4: Betweenness of Crete and Cyprus with respect to the maritime traffic (1283 - 1453)

Conclusions and Future Work

It sounds like a commonplace to describe the Mediterranean Sea, geographically and historically, as an area of intense exchanges and communications; however the fact is that any visualisations up to this point, when they exist, never went beyond the narration and failed to give a concrete idea of the pace imposed by Venetian navigation over a period of 170 years.

With this work, we go beyond that common way of visualising maritime historical data. First, we have designed processing procedures to automatically digitise data present only on paper documents. Second, based on this digitised data, we modelled the Venetian maritime connections over 170 years as a network. Third, we magnified the network over Cyprus and Crete and extracted a measure of betweenness for these two islands.

From a qualitative analysis point of view, we showed the consequences of three historical events with respect to the *Incanto* system. We are confident that we can apply this methodology to better explain historical events and quantify their influence on the global maritime network.

References

1. **Ercole, G.** (2006). *Duri i banchi!: le navi della Serenissima*, 421-1797. Gruppo Modellistico Trentino di Studio e Ricerca Storica.
2. **Lane, F.C.** (1973). *Venice, A Maritime Republic*. ACLS Humanities E-Book. Johns Hopkins University Press.
3. **Luzzatto, G., Padovan G.** (1941). *Navigazione Di Linea e Navigazione Libera, Nelle Grandi Città Marinare Del Medioevo*. Popoli 1: 389-391.
4. **Stöckly, D.** (1995). *Le système de l'Incanto des galées du marché à Venise: fin XIIIe - milieu XVe siècle*. BRILL.
5. Digital Atlas of Roman and Medieval Civilization. darmc.harvard.edu/icb/icb.do (accessed on 30.10.2013)
6. **Bagnall, R., Talbert R. J. A., Elliott T., Twele R., Becker. J., Gillies S., Horne R., McCormick M., Rabinowitz A., and Turner B..** (2006). *Pleiades: A Community-built Gazetteer and Graph of Ancient Places*. Collection. pleiades.stoa.org
7. **Dudan, B.**, (1938), *Il Dominio veneziano di Levante*. Nicola Zanichelli.
8. **Major A.** (1989), *Les colonies continentales de Venise en Grèce méridionale*, XIV-XV siècle, Doctorat Nouveau Régime, Histoire, A.N.R.T..
9. **Balard, Michel.** (1987), *Les Vénitiens En Chypre Dans Les Années 1300*. Byzantinische Forschungen 12: 580–606.

**Canon, value and artistic culture:
critical inquiry about the new
processes of assigning value in the
digital realm**

Rodríguez-Ortega, Nuria

nro@uma.es
University of Málaga, ES

1. Framework of thought and specific purposes

The processes of assigning value to cultural objects, as well as the establishment of the canons which derive from those processes, have constituted until today one of the intellectual, ideological and political foundations of the development of Art History discipline as an institutional discourse (Halbertsma, 2007). This explains why the critical dismantling of the concept of canon as a structure of power, criterion of authority and legitimizing argument represented a significant line of inquiry for the post-structuralism and, more recently, for the postcolonial

theory (Parker and Pollock, 1981; Bloom, 1994; Perry and Cunningham, 1999; Gorak, 2001; Bart, 2005; among others). Especially, it has been emphasized the need to bring out a critical awareness of the multiplicity and heterogeneity that define the processes of assigning value and meaning to objects on the basis of the variety of cultures, genders, races and territories. It has been stated that, in our global world it is essential to understand the concepts of canon and value in terms of plurality and difference. It has also become necessary to explore the specific idiosyncrasies of those processes as a mean to make recognizable and significant that diversity.

Within this framework of thought, the so-called 'digital turn' offers to us another scenario of critical analysis to rethink these issues from the perspective of the new conditions of the digital society, which is modeled by the prevalence of the software, and it is characterized by the potentiality of interactivity, user-generated content, and – at least in theory - global access to and massive distribution of cultural images and objects. This is the intellectual background of my proposal.

As a response to the theme of Digital Humanities 2014 (Digital Cultural Empowerment), I propose to explore how the digital turn, which has brought a new model of society, economy, culture, and a new epistemology (i.e. new ways of production, narration, distribution and consumption of knowledge), is leading to a redefinition of the processes of assigning values -and the values themselves- that have hitherto prevailed in the comprehension of cultural objects within the Art History discipline, resulting in new forms of canonization.

My approach is inspired to the current Digital Humanitie's thought which proposes to rethink the circumstances and the consequences of this 'new' disciplinary field from the perspective of the cultural critique (Lothian and Phillips, 2013; Dacos, 2013; Galina, 2013; Fiormonte, 2012; Liu, 2012; McPherson, 2012; Higgin, 2010, among others). The field of Digital Humanities is becoming aware that there is a real risk of perpetuating in the digital world and in the practice of digital scholarship the same problems of marginality and subalternity that characterized our pre-digital world. In the field of Art History this trend is represented by the super-imposition of specific canons for the understanding and explanation of artistic phenomena. A critical approach to Digital Humanities requires a review of both established and new structures of power that are emerging. However, although the field of artistic culture is one of the most affected by these new processes, the critical discourse is still in its embryonic stage within the context of Digital Art History studies. In my opinion, there is an urgent need to conduct a thorough analysis from the perspective of critical theory. My scope would be to develop such a perspective, unveiling and questioning what kind of art-historical discourses and narratives, and what kind of digital artistic culture we are building on the web (Rodríguez Ortega, 2013).

Now then, we must bear in mind that the building of the digital artistic culture, and the growth of the emerging Digital Art History itself are defined by a dialectical tension between the new processes of assigning value and the maintenance of those traditional structures that had characterized the development of Art History discipline during the Twentieth century [1] (Baca, Helmreich and Rodríguez Ortega, 2013; Kohle, 2013). Examining this tension is a complex task, since these practices and criteria are simultaneously interlaced and in confrontation. Any inquiry must be based then on a dual question: a) we must scrutiny what is really changing in the digital medium in regard to the processes of assigning value to cultural objects, and to what extent these new processes are entailing a destabilization of the traditional criteria of Art History's institutional discourses; in short, the aim is to explore to what extent the Art History discipline and its allied institutions (Museum, Art Criticism, Market, etc.) are being put in crisis as argument of authority; b) perhaps more importantly, we must be aware that, while these changes –sometimes very visible- are taking place, the logic that governs the processes of assigning value based on institutional policies and established power structures is maintained, as well as it is preserved the canons that characterized the critical and conceptual definition of artistic objects and images during the twentieth century – essentially, Western, white and male.

2. Defining hyper-canonicalization and de-canonicalization processes

For this presentation, I will focus on two of theses processes, which are related to the conceptualization of the social web as the new laboratory of cultural production. In this scenario, new actors, hitherto completely unrelated to the traditional ecosystem of Art History (Academy - University, Museums, Critique, Market), arise and perform, fostering a paradoxical redefinition –paradoxical due to its ambivalence- of the traditional concepts of canon and value.

Firstly, I would like to address the process that I propose to call 'hyper-canonicalization' since this type of process superimposes and at the same time encompass the traditional ones. Therefore, as indicated above, a regime based on institutionalism and authorial power structures remains. The challenge lies, then, in determining which are such arising power structures and who has the ability to control them.

In part we can associate this process to the rise of software oligopolies and social networks companies (Google, Facebook, Twitter, Apple, Microsoft, etc.) [2], which belong to the same Western and Anglophone economic-cultural context. They control the technological infrastructures, the algorithms for data processing and retrieving, the channels for content distribution and the social interactions platforms that are used by cultural institutions to interrelate with their audiences (see, for example, the massive presence of museums in social networks as inexcusable part of their communication policies and activities). This indisputable technological and economic supremacy can lead us to new forms of digital colonialism and new cultural monopolies. Some of them are obvious. From my point of view, one of the clearest cases is represented by the Google Art Project, whose declared objective is to become the global gate for accessing the entire collections of museums worldwide. Nevertheless, the philanthropic mission of providing a comprehensive and free access to the objects of world culture underlies the threat that the museum identity can get lost on the web. Each museum, as a differentiated institution, is defined by certain discursive strategies, intellectual positions and critical criteria. However, these signs of identity could dissolve if the collections would be seen preferably 'through' Google. Not surprisingly, it is frequent to find that museums' websites use Google Art Project among their recommended and authorized information sources. Consequently, museums themselves are participating in this process of legitimating Google – the Google Cultural Institute - as a new institutional discourse.

Others are less obvious, but equally disturbing. For example, despite all digital archives and online catalogs developed by public and private institutions, the largest digital images archive and the most accessible is – let admit it - Google Images. Google Images establishes a hierarchy of the images retrieved based on computational procedures that run according to algorithms completely unrelated to the epistemological, aesthetical, historical and/or symbolic specificities of artistic artifacts. Thus, the software, whose conceptualization has nothing to do with these specific aspects, assumes the power of the decision making when 'ordering' the images of our cultural heritage.

'De-canonicalization' is the name that I propose for the second process that I want to address in this presentation. This process emerges directly from the social and distributed users' interactions with the cultural images and objects on the web. Under my perspective, what is in crisis here is the concept of 'canon' itself, because of its bottom-up orientation which dissolves the idea of canon understood as the institutionalization of specific values representing the ideas and interests of those that hold a sort of privileged position of authority (intellectual, economic, political, etc.).

This process is linked to the unprecedented empowerment of social communities to interact with and give new meanings to cultural artifacts through their multiple, heterogeneous, and distributed digital activity. It is thus set up a new scenario that unfolds outside the institutional frame, and whose processes of assigning value are governed by very different criteria [3]. Hence, social memory, subjectivity, emotionality, etc. become fundamental factors for the re-semanticization of cultural objects

and for their relocation in new scales of value. This new context involves a disruption of the principle of authority in the Art History discipline and its allied institutions, which comes into confrontation with these actions in a double way: or ignoring them, or appropriating them.

In fact, the appropriation of the logics of participation and sharing that characterize the web 2.0 is the basis of the so-called 'social museum' (Simon, 2010). Nevertheless, these actions bring about another problematic issue on which we need to reflect critically. Certainly, the valuable social knowledge found in the users' interactions have already been recognized by projects that advocate for a hybrid knowledge (expert plus non-expert), which may result in a new process of assigning value and in a new canonization model. See for example, Your Painting (<http://www.bbc.co.uk/yourpaintings>), a project based on the social tagging of British paintings (Baca, 2013), or the History Harvest (<http://historyharvest.unl.edu>), an open digital archive of historical artifacts collected by various communities through the United States, which are systematized and prepared for research and interpretation by a group of scholars. While recognizing the positive aspects of these initiatives, some questions arise To what extent the institutions are appropriating these logics of participation and sharing in order to subsuming them as part of their institutional discourses and canons? To what extent are we facing a phenomenon of 'domestication' and an attempt to attract the outsiders to the 'center', establishing a sort of 'controlled' framework for their activities, such as perturbing sometimes for the institutions?

3. Open questions: What are facing?

I will conclude with a set of open questions that underlie this approach and that should be discussed in depth in following studies: To what extent the Art History discipline is possible outside an institutional framework? Is that condition an argument to explain the need for operating an institutionalization of the digital environment, which is, by nature, open, distributed, and multiple? To what extent the discipline of Art History and its allied institutions are willing to share their position of authority, at least consciously? And to what extent they are aware that they are yielding this position to new structures of power? Recently, James Cuno (President and CEO of the Getty Trust) wondered from a postcolonial perspective: Who owns the past? (Cuno, 2013). Now, I think, it is the time to ask: Who owns the value and the canon in the digital realm? Who has the ability to assign value to cultural objects and images? Who holds now the authority and power to establish the new canons and legitimizing discourses in the context of digital society?

Notes

[1] We should not forget that the dialectic tensions and contradictions have been defining factors in the development of the Art History discipline since its early beginnings (Donald Preziosi. *Rethinking Art History*. Yale University Press, 1989). Therefore, the challenge now is to examine which are the new factors that participate in this process.

[2] Regarding the new inclusion-exclusion regimes associated to the software oligopolies, see Juan Martín Prada. *Prácticas artísticas e Internet en la época de las redes sociales*, Madrid: Akal, 2012.

[3] As examples, see the following projects: www.Bodebarna.net; or www.cabanyalarchivovivo.es/index.html. Both initiatives are based on the appropriation by social communities, belonging to a specific territory, of the cultural heritage related to such territory, using for that digital infrastructures and strategies. The objective is to give them – both cultural heritage and territory- new meaning and value, and rethinking them from the point of view of the social memory and collective interests.

References

Baca, Murtha, Anne Helmreich and Nuria Rodríguez Ortega (2013). *Digital Art History*. Special double issue of Visual Resources. An international Journal of Documentation. vol. XXIX (1-2), march-june.

Baca, Murtha (2013). "The Public Catalogue Foundation's Your Panting Project", Visual Resources, XXXIX (3), 151-153

Bart, J. M. van der Aa (2005). *Preserving the Heritage of Humanity? Obtaining World Heritage Status and the Impacts of Listing*. Groningen: University Library Groningen.

Bloom, Harold (1994). *The Western Canon: The Books and School of the Ages*. New York: Harcourt Brace.

Cuno, James (2013). "Who Owns the Past? Encyclopedic Museums in the Post-Colonial Present", in Luis Arciniega, ed., *Memoria y Significado. Uso y recepción de los vestigios del pasado*, Valencia: Universidad de Valencia, 215-227

Dacos, Marin (2013). "La stratégie du Sauna finlandais", in Blogo Numericus, Mayo 2013. Disponible en: <http://blog.homo-numericus.net/article11138.html> [octubre 2013].

Fiormonte, Domenico (2012). "Towards a Cultural Critique of Digital Humanities", Historical Social Research – Historische Sozialforschung, Special Issue, no. 141, HSR vol.37 (2), 59-76. Disponible en: http://www.cceh.uni-koeln.de/files/Fiormonte_final.pdf [octubre 2013].

Higgin, Tanner (2010). "Cultural Politics, Critique and Digital Humanities", in Gaming the System, 25 Mayo de 2010. Disponible en <http://www.tannerhiggin.com/cultural-politics-critique-and-the-digital-humanities/> [octubre 2013].

Galina, Isabel (2013). *Is There Anybody Out There? Building a global Digital Humanities community*. Keynote speech. Digital Humanities Annual Conference, Nebraska (USA). Disponible en: <http://humanidadesdigitales.net/blog/2013/07/19/is-there-anybody-out-there-building-a-global-digital-humanities-community/> [octubre de 2013].

Gorak, Jan, ed. (2001). *Canon vs Culture: Reflections on the Current Debate*. New York: Garland.

Halbertsma, Marlite (2007). "The call of the canon. Why art history cannot do without". In Elizabeth Mansfield, ed., *Making Art History. A changing discipline and its institutions*, New York and London: Routledge, 16-30

Kohle, Hubertus (2013). *Digitale Bildwissenschaft*. Glückstadt: Verlag Werner Hülsbusch.

Lothian, Alexis and Amanda Phillips (2013), "Can Digital Humanities Mean Transformative Critique", *Journal of E-Media Studies* 3.1.

Liu, Alan (2012). "What is Cultural Criticism in the Digital Humanities?", in Mathew K. Gold, ed., *Debates in the Digital Humanities*, University of Minnesota Press. Disponible en: <http://dhdebates.gc.cuny.edu/debates/text/20> [octubre 2013].

McPherson, Tara (2012). "Why are the Digital Humanities so White?", in Mathew K. Gold, ed., *Debates in the Digital Humanities*, University of Minnesota Press, 2012. Disponible en: <http://dhdebates.gc.cuny.edu/debates/text/20> [octubre 2013].

Parker, Roszika and Griselda Pollock (1981). *Old Mistresses: Women, Art and Ideology*. New York: Pantheon.

Perry, Gill and Colin Cunningham (1999). *Academies, Museums, and Canons of Art*. New Haven, CT: Yale University Press in Association with Open University.

Rheingold, Howard (2002). *Smart Mobs. The Next Social Revolution*, Cambridge: Basic Books

Rodríguez Ortega, Nuria (2011). «Narrativas y discursos digitales desde la perspectiva de la museología crítica», Museo y Territorio, Fundación General de la Universidad de Málaga, n. 4, 14-29

Rodríguez Ortega, Nuria (2013). "Digital Art History: An Examination of Conscience", Digital Art History. Special double issue Visual Resources. An international Journal of Documentation. Edited by Murtha Baca y Anne Helmreich and Nuria Rodríguez Ortegavol. Vol. XXIX (1-2), march-june 2013, 129-133.

Simon, Nina (2010). *The Participatory Museum*, Museums 2.0, Santa Cruz.

Discourses and Disciplines in the Enlightenment: Topic Modeling the French Encyclopédie

Roe, Glenn

glenn.roe@anu.edu.au

The Australian National University

Gladstone, Clovis

clovis.gladstone@uchicago.edu

The University of Chicago

Morrissey, Robert

rmoriss@uchicago.edu

The University of Chicago

1. Introduction

Discourse analysis has for the past half century been a staple of the text-based historical and social sciences. Part and parcel of the 'linguistic turn' of the 1960s, French discourse analysis was furthermore one of the first disciplines to embrace computational text processing when Michel Pécheux developed a computer program to identify ideological processes in textual corpora¹. Steeped in contemporary linguistic theory, Pécheux and his team sought an automated method for uncovering hidden ideological meanings in text corpora. That same year, Michel Foucault's *L'Archéologie du savoir* broadened the conception of 'discourse' and of the underlying power politics at play in its formation². Diverging significantly with Pécheux, Foucault's analytical model of 'archaeology' brought with it a less strictly-linguistic approach to the discursive. By "loosening the embrace [...] of words and things", discourses are understood "as practices that systematically form the objects of which they speak"³.

This expanded notion of discourse would go on to exert a strong influence on French historical studies, and in particular, on the historiography of the French Enlightenment and Revolutionary periods, as is evident in the work of François Furet, Lynn Hunt or Keith Baker^{4 5 6}. More recently, Sophia Rosenfeld and Dan Edelstein have re-introduced the specifically linguistic elements of discourse analysis back into the historian's and literary scholar's toolbox, most notably through the analytical use of historical and linguistic databases^{7 8}.

With the rapid growth of digital text collections, a revisiting of Pécheux's earlier notion of an 'Automatic Discourse Analysis' approach would seem warranted, particularly given recent developments in information retrieval such as topic modeling⁹. This paper is thus an attempt at reconciling the computational and the discursive, using topic modeling to uncover Enlightenment discourses in the *Encyclopédie* of Diderot and d'Alembert. Moreover, Foucault's concept of archaeology is used to justify topic modeling's 'bag of words' analytical model, as it frees us from exclusive interest in language structure, and what that structure conveys, and orients us more towards the association of the various ideas or 'topics' that form a discourse.

2. Topic Modeling as Discourse Analysis

Topic modeling is a machine learning approach that was originally designed as a way to classify large amounts of text with minimal human intervention¹⁰. David Newman and Sharon Block have furthermore demonstrated through their use of pLSA (Probabilistic Latent Semantic Analysis) that such unsupervised algorithms can provide a unique overall picture of the contents of a corpus by organizing the data in a manner that gives researchers an objective and wholly original perspective on the texts being analyzed¹¹. Of course, categorizing texts with no human interaction is not something humanists can accept without question, and this critical point forms the basis of our

previous experiments with supervised classification algorithms such as Naive Bayes and Vector Space^{12 13}.

For this project, we employ the Latent Dirichlet Allocation (LDA) algorithm as it is built upon the important premise that documents, however focused, are never about one single topic, but are the result of multiple topics bound together in a single text unit¹⁴. Consequently, the documents analyzed by this algorithm will be identified by a unique signature: a distribution of topics that represents the variety of things discussed in them. As these clusters of words do not necessarily map onto what humanists consider a 'topic' or theme, we judge their coherence based less on their thematic consistency and more as the representation of a particular discourse, where closely related concepts are used together in a given context. One could imagine, for instance, finding a discourse that never seems to be the main subject of any document, but that nevertheless runs through a significant number of them.

3. Use Case: Topic Modeling the Encyclopédie

As a use case, we have chosen to examine one the Enlightenment's exemplary texts, the *Encyclopédie* of Diderot and d'Alembert¹⁵. Our aim here is to use LDA to go beyond the disciplinary boundaries of the editors' original classification scheme, which was designed (along with the cross-references) to connect articles amongst themselves across the whole work, but in reality did little to provide guidance to its readers. The physical structure of the text, which caused articles to be read in relative isolation from others, made obtaining a full dialogic perspective of any given class or article unrealistic. By using topic modeling as a discourse analysis tool we aim to highlight each article's unique discursive makeup. This will allow us to generate a more transversal view of the encyclopaedia and its contents. Whereas David Blei has asked: "What is the likely hidden topical structure that generated my observed documents?"¹⁶; we likewise ask, "what are the non-obvious discourses that span across multiple disciplines in the *Encyclopédie*?"

From a methodological perspective, we are using the well-known machine learning toolkit MALLET with several Python wrap-arounds¹⁷. Since our goal is to uncover discourses across the entire *Encyclopédie*, we settled on a relatively low number of topics (between 280 and 360) compared to the total number of classes of knowledge (2,900), but this number was consistent with our previous machine classification experiments¹⁸. Once our topic model was generated, we stored the results in a SQLite table, along with all available metadata. We then wrote a web interface to visualize this database and run queries against the original metadata.

Using the above interface we were able to identify many of the *Encyclopédie*'s disciplinary vocabularies in our topic lists, which were verified using article metadata. Not surprisingly, the 'chemistry' topic was found most in chemistry articles, the 'botany' topic in botanical articles, mathematics in mathematics, etc. What interests us, however, are topics that are both distinct in nature -- i.e., identifiable with a particular 'discourse' -- and that span multiple disciplinary boundaries. Mapping these discourses through the various classes and articles in which they are prevalent leads to a greater understanding of the dialogic and discursive elements at play in the seemingly innocuous encyclopaedic classification system.

The topic we have identified with the discourse 'droit naturel', for instance, is present in more than 60 grammar articles, almost double that of its own class (Figure 1).

Topic #56 'droit naturel': droit lois nature société loi hommes raison choses état homme justice naturel naturelle juste vie gens devoirs morale vertu souverain...

Article classes:

unclassified, 191

Grammaire, 61

Jurisprudence, 56

Morale, 55

Droit naturel, 30

Géographie moderne, 28

Théologie, 27

Géographie, 25

Droit politique, 23

Histoire moderne, 22

Articles in 'Grammaire' with topic weight:

Diderot, INDISPENSABLE, 0.6273819846678368

Jaucourt, RÈGLE, RÉGLEMENT, 0.6010994508849054

Diderot, CONDUITE, Grammaire, 0.5303738356879001

unknown, INOBSERVANCE, ou INOBSEUR, 0.5288916753532583

unknown, MAXIMES, Grammaire, 0.5058774638177999

Diderot3, INVOLABLE, Grammaire, 0.4695993200405884

d'Alembert, CONTRAINdre, OBLIGER, FORCER, 0.4533961077730776

Diderot, INIQUE, INIQUITÉ, Grammaire, 0.4400953539002383

Diderot, Bien, (homme de) homme d'honneur, 0.4238494003569378

Diderot3, SUPPLANTER, 0.3929385730821571

Fig. 1: Topic #56: "Droit Naturel"

Among the top grammar articles we find the small unsigned article 'Inviolable' that has since been attributed to Diderot. In it, alongside the grammatical definition of the term, we find a usage example that reads: "La liberté de conscience est un privilège inviolable" (8:864) -- a reference that subtly places freedom of thought amongst other 'natural' and unalienable rights. We find a similar treatment in the article 'Supplanter', which contains a thinly-veiled condemnation of tyranny as an unnatural state of governance.

Other classes function in much the same way as Grammar, allowing the *philosophes* to smuggle controversial opinions into articles of a seemingly neutral scope. By tracing the presence of various discourses in an inter-disciplinary manner we can begin to uncover the various subversive, discursive, and ideological practices in play over the entirety of the *Encyclopédie*. The discourse around morality, for instance, is found in no less than 94 articles from the 'géographie ancienne' class (Figure 2).

Topic #227 'morale': homme esprit hommes amour vertu morale notre caractere coeur ame mal meurs raison passions bonheur société vice choses plaisir nos...

Article classes:	Articles in 'Géographie ancienne' with topic weight:
Grammaire, 532	Diderot, ARBELLE, 0.18277249239163076
unclassified, 414	Diderot, DRANSES, 0.16505879453827525
Moral, 220	unknown, LIGURIENS, Ligurini, 0.1459712481859705
Géographie moderne, 146	Jaucourt, SELINUITE en Cilicie, 0.1382502735986093
Géographie ancienne, 94	Jaucourt, VOLATERRAE, 0.13348542137754754
Histoire moderne, 91	Jaucourt, SALMACIS, 0.12328095063070527
Mythologie, 90	Jaucourt, SELEMNUS, 0.12271914960203098
Géographie, 83	Jaucourt, MOSYNIENS ou MOSYNOEICIENS, 0.10255446055404507
Synonymes, 77	Jaucourt, ULUBRAE, 0.10180768346042847
Jurisprudence, 67	unknown, TRINEMEIS, 0.10123411273926909

Fig. 2: Topic #227: "Morale"

Diderot is here again exemplary in his discursive acrobatics. Whilst describing a tribe of ancient Thracians in the article 'Dranses', he quickly turns the discussion towards moral relativism (in a move the prefigures his later work, *Le Supplément au voyage de Bougainville*), with the assertion that: "Ce n'est pas la nature, c'est la tyrannie qui impose sur la tête des hommes un poids qui les fait gémir & détester leur condition" (5:106).

A similar deployment of the discourse around 'le culte religieux' -- a subject on which the *encyclopedistes* were forced to tread lightly -- can be found in the more than 100 articles labeled as 'histoire moderne' (Figure 3).

Topic #242 'culte religieux': religion dieu hommes culte dieux chrétiens ciel eux divinité christianisme terre monde esprit superstition vie homme payens doctrine mystères opinions...

Article classes:	Articles in 'Histoire moderne' with topic weight:
unclassified, 177	Diderot2, SCHOOUBIAK, 0.49114822022547194
Théologie, 140	d'Holbach5, ZENDICISME, 0.44992922706317456
Histoire moderne, 110	Jaucourt, DRAGONADE, 0.38727007830542115
Histoire ecclésiastique, 104	d'Holbach5, PASENDA, 0.36537912347852947
Grammaire, 89	d'Holbach5, XENXUS, 0.359198844836053
Géographie moderne, 76	Mallet, CHUPMESSATHITES, 0.3509211906764424
Mythologie, 75	d'Holbach5, SHECTEA ou CHECTEA, 0.34768525234924785
Critique sacrée, 50	d'Holbach5, RITES, 0.3429275926917365
Géographie, 49	Diderot, HOURIS, 0.32068644910614613
Géographie ancienne, 49	d'Holbach5, OSSA-POLLA-MAUPS, 0.290310143692191

Fig. 3: Topic #242: "Culte Religieux"

The article 'Schooubiak', for instance, is another unsigned (but later attributed) article by Diderot, in which he describes an Islamic sect that practices an unusual form of religious tolerance. This seeming incongruence with the accepted cultural stereotypes of the time allows Diderot to raise the issue of religious intolerance thus using the sect as a proxy for the *philosophes* themselves, and condemning those who would oppose them: "les prêtres étant les mêmes par-tout, il faut que la tolérance soit détestée par-tout" (14:778).

As we have seen above, many of these encyclopaedic discourses were deployed subversively in order to move the narrative of Enlightenment forward, and indeed, the discursive nature of the Enlightenment in France has recently been brought to light¹⁹. By extending the reach of our topic modeling approach to other Enlightenment texts we can begin to identify the discursive practices of texts and authors on an even greater scale and with a greater level of systematicity. As a philosophic war machine, as well as a contemporary reference work, the *Encyclopédie* is an ideal starting point for this sort of work.

Many of the discourses we find therein may have been lost to the modern reader through the classification process itself, and still others may prove useful in uncovering interdisciplinary connections that would have otherwise gone unnoticed.

References

1. Michel Pêcheux (1969), *Analyse automatique du discours*, Paris: Dunod.
2. Michel Foucault (1989), *L'Archéologie du savoir*, Paris: Gallimard, (1969) [English translation: The Archaeology of Knowledge, London: Routledge].
3. Foucault (1989), p. 54.
4. François Furet (1978), *Penser la Révolution française*, Paris: Gallimard.
5. Lynn Hunt (1984), *Politics, Culture and Class in the French Revolution* Berkeley: University of California Press.
6. Keith Michael Baker (1990), *Inventing the French Revolution: Essays on French Political Culture in the Eighteenth Century* Cambridge: Cambridge University Press.
7. Sophia Rosenfeld (2001), *A Revolution in Language: The Problems of Signs in Late Eighteenth-Century France*, Stanford: Stanford University Press.
8. Dan Edelstein (2009), *The Terror of Natural Right: Republicanism, the Cult of Nature, and the French Revolution* Chicago: University of Chicago Press.
9. For a thorough introduction and discussion of topic modeling for humanistic research, see the special issue of the *Journal of Digital Humanities* edited by Scott Weingart and Elijah Meeks (2012): Journal of Digital Humanities Vol. 2, No. 1 Winter [journalofdigitalhumanities.org/2-1].
10. David M. Blei, Andrew Y. Ng, and Michael I. Jordan (2003), *Latent Dirichlet Allocation*, Journal of Machine Learning Research 3 (4–5): 993–1022.
11. David J. Newman and Sharon Block (2006), *Probabilistic Topic Decomposition of an Eighteenth-Century American Newspaper*, Journal of the American Society for Information Science and Technology 57(6): 753–67.
12. Russell Horton, Robert Morrissey, Mark Olsen, Glenn Roe, and Robert Voyer (2009), *Mining Eighteenth Century Ontologies: Machine Learning and Knowledge Classification in the Encyclopédie*, in: Digital Humanities Quarterly 3.2 Spring.
13. Charles Cooney, Russell Horton, Mark Olsen, Glenn Roe, and Robert Voyer (2008), *Hidden Roads and Twisted Paths: Intertextual Discovery using Clusters, Classifications, and Similarities*, Digital Humanities 2008, eds. Lisa Opas-Hänninen, Mikko Jokelainen, Ikka Juuso, and Tapio Seppänen, University of Oulu, Finland: 93–4.
14. Blei et al. (2003).
15. All data and references are drawn from the ARTFL digital edition of the *Encyclopédie*, developed by the University of Chicago's ARTFL Project: [encyclopedia.uchicago.edu].
16. David Blei (2013), *Topic Modeling and Digital Humanities*, Journal of Digital Humanities 2.1.
17. See [mallet.cs.umass.edu/].
18. Horton et al. (2009).
19. Dan Edelstein (2010), *The Enlightenment: A Genealogy* Chicago: University of Chicago Press.

National Data Curation and Service Center for Digital Research Data in the Humanities

Rosenthaler, Lukas

lukas.rosenthaler@unibas.ch
Digital Humanities Lab / University of Basel

Fornaro, Peter

peter.fornaro@unibas.ch
Digital Humanities Lab / University of Basel

Clivaz, Claire

claire.clivaz@unil.ch
LADHUL / University of Lausanne

For quite a few years the long-term preservation of digital data and resources has been an ongoing topic within the IT-industry and archiving community. While the OAIS reference model offers a very reasonable framework for long-term archival of digital data such as digitized images, sound or text documents, the archival of highly structured digital data such as databases still poses a lot of problems. Flattening databases to XML text files has been used successfully to archive the contents of relational databases (RDBMS) ^{1 2 3}. However this method reduces the accessibility, since the XML-files have usually to be read back into a RDBMS to be used. Today's best practice to keep the usability of structured data as high as possible is to migrate data repositories and its software environment (user interfaces, analytical tools etc.) to new technology to ensure its accessibility ⁴. Yet, replacing obsolete hardware and software infrastructure is an ongoing labor-intensive process that requires continuous financial effort. In addition, given that online research data is usually constantly being modified to reflect new findings and thus is changing dynamically, referencing it (e.g. for citations) is not straight forward.

Despite these difficulties, the use of digital research data including databases has become very common in humanities. At the same time, the term itself of "data" is not sufficient to describe the resources used and produced by the humanist researchers: collections of digital data, digitized manuscripts, collections of digitized photographs and metadata related to it, as well as new digital resources and objects produced in the digital cultural framework. As long as project funding is available, many of these digital sources are made available to the research community. However, after the funding ceases, most of these digital sources will remain accessible only as long as the hardware and software remains in working order. After some time – typically some years – most data will go offline because of lack of maintenance. Thus, most of the digital sources created within research projects will have a rather short lifetime and are no longer available for the research community after some time. However, these digital sources are a valuable base for possible new projects but its sustainability is not ensured due to missing fundings.

Given this disappointing situation, the Swiss Academy of Humanities and Social Sciences (SAHSS) - on the behalf of the State Secretariat for Education, Research and Innovation (SERI) - has launched a project to address this situation in the national context of Switzerland. The Digital Humanities Lab of the University of Basel (DHLab) in conjunction with the Universities of Lausanne (Ladhu) and Bern, in association with the Swiss National Archives, participated in a tender and have received the task to establish a solution. In a first two years period, a pilot for a "National curation and service center for digital data in Humanities" (DCSC) will be established. Using several test cases of different sizes and complexity from different disciplines, the methods and processes, legal aspects, infrastructure needs, and last but not least, the cost and expenses, have to be evaluated. The proposed DCSC is based on the following premises:

- Preserving software in a useable and working condition is still a very difficult task, as illustrated by the recent meeting "Preserving.exe: Toward a National Strategy for Preserving Software" by the American Library of Congress ⁵. Thus it would be difficult and costly to maintain a multitude of different systems for a long time. Emulation of obsolete hard- and software as proposed by R. A Lorie ⁶ is also very difficult and has its share of problems (for example see ⁷). Therefore the different digital sources or databases have to be integrated into a minimal number of hard- and software infrastructures. Ideally, only one hard- and software system has to be maintained.
- In a first phase the adoption of existing data sources of research projects at or beyond the end of funding will be dominant. In a second step the goal must be that researchers of ongoing or future research projects are escorted through

the creation and use of digital sources in order to facilitate the accessibility of the data after the end of funding.

- For the exchange with other platforms and infrastructures, the DCSC implements interfaces for import, export and querying of information (as far as not restricted by legal constraints such as copyright issues and/or protection of personal rights). The adopted digital sources must be accessible through a powerful user interface for search/analysis, a RESTful web-service or as SPARQL-endpoint in order to integrate the data into other research projects and/or databases.
- The DCSC should encourage new research models in order to allow for optimal use of digital sources and to propose efficient training modules and support for all the new research projects funded by the Swiss National Science Foundation.
- International contacts are a key-point for this center, in order to prepare Swiss digital Humanities research to be interrelated to international research.

Given the nature of research in humanities, we expect data sources the DCSC has to deal with to be very heterogeneous and consisting mostly of qualitative data (which is possibly linked to digitized objects). We have chosen to use the virtual research environment SALSAH ^{8 9} as a technical platform for consolidation of different data sources. SALSAH is RDF/RDFS-based and thus well suited to emulate the basic functionality of RDBMS's, simple databases such as MS-Access, FileMaker etc. SALSAH is currently actively developed by the DHLab.

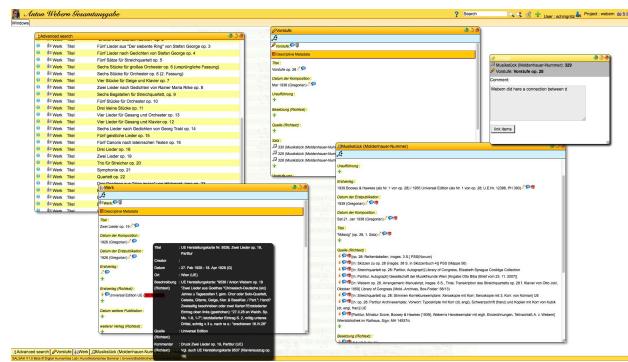


Fig. 1: The webinterface of SALSAH implements a desktop metaphor within the webbrowser window in order to work with multiple sources simultaneously.

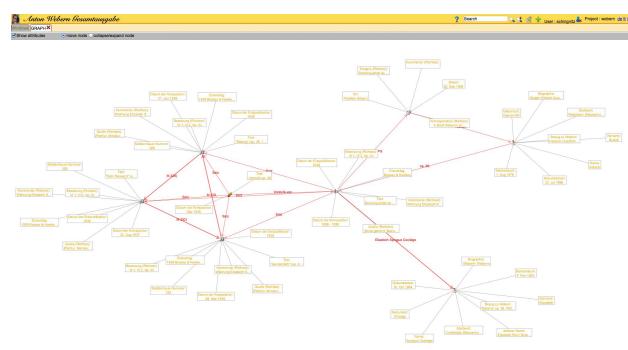


Fig. 2: SALSAH allows a dynamic visualization of the graph-like structure of the RDF-representation of the data.

Within a research project funded by the Swiss National Science Foundation SALSAH is currently being extended with several new important features (expected in 2015):

- a "time machine" which will allow digital objects to be referenced by permalinks which include the time of referencing. Thus such a permalink will always show the digital object in the state it had at the time being referenced. These permalinks will add true "citetability" to the SALSAH environment.
- SALSAH, which is currently organized as a (technically) centralized system, will be transformed into distributed, self-organizing P2P system. At the same time, an archival system based on DISTARNET ¹⁰ will be added to SALSAH

to secure the against data loss due to catastrophic events like hardware failure, flooding, fire etc. at any SALSAH location. DISTARENTE also uses P2P technology to maintain redundant multiple backups of the data within the network.

- Within the DCSC project, SALSAH will be extended to support "open data"-standards¹¹ for access and "linked data"¹². However, open access may be restricted by legal reasons (copyright, privacy etc.). SALSAH includes a fine-grained identity and rights management.

SALSAH will be continuously enhanced according to the needs of the researchers using the platform. It is planned to move SALSAH to "open source" by the end of 2014.

The main tasks of the DCSC will be threefold:

1. Maintaining the technological infrastructure and adapting it to the needs of the researchers and changing technology. This task will be located in Basel during the pilot phase, but since SALSAH will become an open source project, other institutions and individuals may contribute to the SALSAH base. However, in our experience open source projects need a powerful "coordinating institution" in order to be successful. The DHLab will be available to play this role.
2. Assistance to the researchers. In Humanities many researchers working with digital sources do not have the technical knowledge to fully exploit the advantages of the digital processes. The DCSC will support the researchers to use digital methods and tools in the best possible way for their research, and encouraging training and education.
3. To create a report with recommendations on how to proceed and transfer the pilot project into a permanent institution.

It has to be noted that the projects funding goes way beyond a "normal" scientific project funding and shows the strong commitment of all involved parties (SAHSS, SERI, Swiss National Science Foundation) to define a persistent national data curation and service center for digital research data in the humanities. The team composed of the University of Basel, Bern and Lausanne and the National Archives demonstrates that as well. The project pilot phase is financed by the SAHSS and the SERI. Since Switzerland is a multilingual nation with a highly federalistic structure, we decided to create the DCSC as a "virtual" center where the technological infrastructure currently will be located and maintained in Basel, but all the other tasks will be performed by local "branch offices" which are very close to the researches; during the pilot step, Lausanne and Bern are testing this "branch office" or "satellite" model. As soon as SALSAH has the P2P functionality implemented, also the technical infrastructure may be distributed if necessary or desired. Given the DISTARNET archival system, the data is secured against loss without the necessity of the local branches to build an expensive and complicated backup infrastructure.

References

1. SIARD of the National Archives of Switzerland, www.bar.admin.ch/dienstleistungen/00823/00825/index.html?lang=en
2. XENA of the National Archives of Australia, xena.sourceforge.net
3. kopal/koLibRI of the german Kopal-project, kopal.langzeitarchivierung.de/index_koLibRI.php.en
4. **The Consultative Committee for Space Data Systems** (2012): *Reference Model for an Open Archival Information System (OAIS) - recommended Practice*, CCSDS Secretariat, Space Communication and Navigation Office, 7L70, Space Operation Mission Directorate, NASA Headquarters, Washington, DC 20546-001, USA
5. **Preserving.exe: Toward a National Strategy for Preserving Software**, May 20-21, 2013, Library of Congress, Washington DC, see also www.digitalpreservation.gov/meetings/documents/othermeetings/preservingsoftware2013/Preserving_Exe_Agenda.pdf
6. **Lorie R. A.** (2001). *Long term preservation of digital information*. Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries, Roanoke, Virginia, United States. 24–28 June 2001. New York, NY: Association of Computing Machinery. pp. 346-352 doi:10.1145/379437.379726

7. **Feng Luan, Mads Nygard, Thomas Mesti** (2010), *A survey of digital preservation strategies*, World Digital Libraries, Vol3 (2), IOS Press

8. See www.salsah.org for the generic entry point, www.salsah.org/dokubib for an (simplified) entrypoint for the Documentation Library of St. Moritz, and www.salsah.org/kuhaba for the Kunsthalle Basel.

9. **Lukas Rosenthaler** (2012), *Virtual Research Environments. A New Approach for Dealing with Digitized Sources in Research in Arts and Humanities*, in: Claire Clivaz u.a. (HG.): "Reading Tomorrow. From Ancient Manuscripts to the Digital Era", Lausanne 2012, S. 661-670, Ebook on www.ppur.info/lire-demain.html

10. **Ivan Subotic, Lukas Rosenthaler and Heiko Schultdt**, *A Distributed Archival Network for Process-Oriented Autonomic Long-Term Digital Preservation*, ACM Proceedings of the Joint Conference On Digital Libraries, ACM New York (2013), pp 29-38

11. See Open Knowledge Foundation, okfn.org

12. See linkeddata.org

Mixed data, mixed audience: building a flexible platform for the Visionary Cross project

Rosselli Del Turco, Roberto

roberto.rossellidelturco@gmail.com

Università di Torino

1. Introduction

Due to an exceptional event, a fragment of *The Dream of the Rood* poem is found on the Ruthwell Cross (Dumfriesshire, Scotland) in the form of an inscription in runic characters; another fragment of the same work, although much smaller, is visible on the Brussels Cross (St. Michael and St. Gudula Cathedral, Brussels, Belgium); the two monumental crosses of Ruthwell and Bewcastle (Cumbria, England) share the presence of runic inscriptions and the same type of carved decorations; the two poems *Elene* and *The Dream of the Rood* (full version) are part of a *florilegium* of religious works, the Vercelli Book (MS CXVII, Biblioteca Capitolare di Vercelli, Italy), in virtue of the central role that the Cross plays in both: the first poem belongs to the texts inspired by the legend of the *inventio crucis* by St. Elena, mother of emperor Constantine, while in the second one the Cross itself appears in dream to the author and tells its own story. These witnesses of the Anglo-Saxon Middle Ages are closely related from a thematic point of view as well as for their contents¹: one could say that this specific thematic cluster was handled by means of a sophisticated multimedia approach by Anglo-Saxon authors, especially visible in the case of the Ruthwell and Bewcastle crosses. The Visionary Cross project² aims at creating a mixed media edition of these artifacts putting together not only the critical edition of the poem, which is going to be showed together with the digitized images of the Vercelli Book, but also the three-dimensional data related to the Crosses.

2. Methodological issues I: integration of heterogeneous data

Within the scope of the project I am currently working with the ISTI-CNR researchers (Pisa) to refine and improve the current version of a combined 3D model / textual edition browsing software³, and to prepare a digital edition of the runic version of the *Dream of the Rood* poem. During this work we faced two critical issues:

- the integration of heterogeneous data on the web platform that we are building to visualize the edition: namely, the 3D

model of the Ruthwell Cross, which is in a standard format (PLY - Polygon File Format⁴) but needs specialized software libraries to be displayed in a Web browser window; the digitized images of the Vercelli Book manuscript; and finally, the edition texts (diplomatic and critical editions of the poem fragment);

- the adoption of an encoding standard that would allow to preserve the different layers of the text (on the graphic level: rune characters vs. transliterated characters vs. modern rendering of text; on the edition level: diplomatic, interpreted and critical editions) at the same time connecting seamlessly with the 3D and 2D elements included in the web platform.

The two issues mentioned above are currently being solved by means of a software framework designed to be flexible enough to be adapted to different types of media and to cope with the specific characteristics of each object; and by use of the TEI XML schemas⁵, in a *ad hoc* customization using only the TEI P5 modules needed for our purposes, to accomplish an encoding of the runic text which is fully interoperable with both the web-based visualization system and the tools commonly used in the digital philology field. The rationale being that, while 3D models have recently started to appear and gain in popularity, they still are self-contained objects with little to no concession to the (possibly very relevant) textual content of the original artifact: our goal is to integrate the separate components in such a way that all the subtle interrelations and connections between the different objects, in short their original multimedia nature, are exposed and made explorable in a flexible browsing environment. For instance, the runes are inscribed on the two narrower sides of the Ruthwell Cross and it is likely that each side, including the larger ones with figural panels, were to be "interpreted" in turn according to the movement of the sun⁶; furthermore, the Bewcastle Cross shows similar iconographic content, but with subtle differences that can be ascribed to the influence of larger cultural traditions (Celtic, Roman) on the work of local artisans.

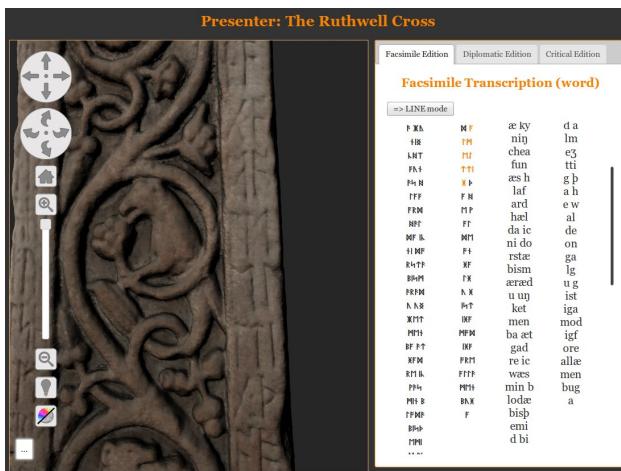


Fig. 1: Experimental edition of the runic text of the Ruthwell Cross

3. Methodological issues II: defining the user and what s/he can do with the edition

While exploring the future shape of the aforementioned environment, though, we came up with new research questions begging for an answer:

- who is going to be the "typical" user of our edition? This is apparently an easy question, but when we considered possible use cases we came up with many more than first anticipated, so that the perspective audience almost looks as heterogeneous as the different media of the edition; taking into account these use cases has led to new functionality being added to the browsing environment, and to the decision that it should be as modular and flexible as possible for future expansion;
- what is our user going to do with our edition? Another apparently innocuous question, especially since we

considered from the start, besides simple browsing of the content, the possibility of user annotation of edition objects; what is different, as we could try for ourselves, is that giving more powers to the user leads to interesting (and potentially risky) new scenarios.

A traditional edition lets the user verify the soundness of the editor's *constitutio textus* by means of the critical apparatus: the latter is a very successful compromise dictated by the limits of the printed edition format, but a digital edition can go beyond such limits and make it possible not only to verify the current version of the edited text but also to experiment with alternatives⁷, creating a "personal edition" on the basis of the available material and the tools provided by the browsing platform, and sharing it according to the "social edition" concept⁸, although this will be limited to the "collaborative annotation" feature described by Siemens rather than to the "user derived content" one (in other words, we are not thinking of a full "collaborative edition"). Note that in theory this is possible not only with regard to textual content, but also for the 3D reconstruction we offer: the Ruthwell Cross is suffering by problems of incompleteness (the horizontal arm is a spurious piece dating to the XIX century), wrong reconstruction (the top piece was placed back-to-front), legibility of the inscriptions (part of the runes have been effaced by the prolonged exposure to the weather and other damages): making the 3D model dynamic and open to alternative positioning of selected parts, and offering digital restoration tools to insert "digital runes" where the original ones are no longer visible, it will be possible to build and check different theories about the Cross and the text it bears.

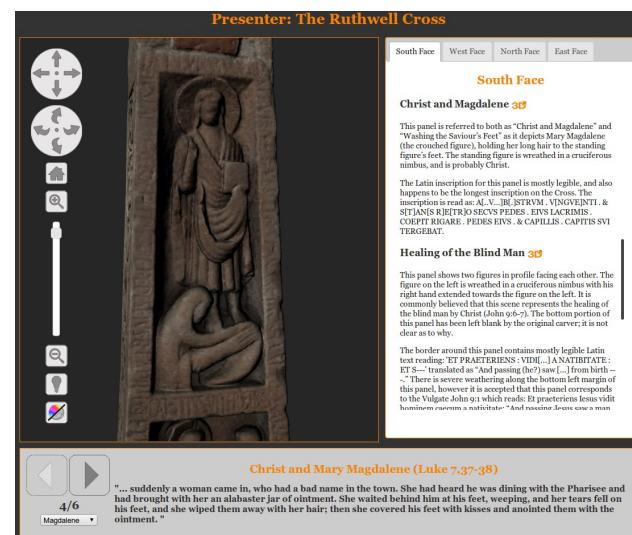


Fig. 2: The Ruthwell Cross 3D model as a teaching tool, clicking on the presentation arrows the model rotates and moves to the location described in the slide

We intend to offer suitable tools to verify our own (as editors) hypotheses, but also to create and manage new hypotheses and to share them with other users, be they related to the textual layers of the edition or to the material aspects of the artifact(s) available in the browsing environment. To do this we will have to go beyond a simple catalog of possible use cases, but also prepare the environment for different user "roles", assigning to each role an appropriate set of capabilities (and responsibilities). While the original plan was that of a single environment for all types of possible users, in fact, our initial work with text and 3D models convinced us that it is impossible to conflate together extremely heterogeneous features. Hence the need to set up a very small number of different environments in which part of the functionality is shared and always present, while other features are specifically targeted to a particular type of user.

Conclusion

This paper will report on the methodological issues described above, explaining which solutions have already been found and at least in part implemented in the browsing environment, which issues are still open and how the project researchers intend to deal with them.

References

1. Karkov, C., Keefer, S.L. and Jolly, S.L. (eds.) (2008). *Cross and Culture in Anglo-Saxon England*. Morgantown: West Virginia University Press.
2. Project web site: www.visionarycross.org/.
3. Callieri, M., Leoni, C., Dellepiane, M., Scopigno, R. (2013). *Artworks narrating a story: a modular framework for the integrated presentation of three-dimensional and textual contents*. ACM WEB3D - 18th International Conference on 3D Web Technology, pp. 167-175.
4. See [en.wikipedia.org/wiki/PLY_\(file_format\)](http://en.wikipedia.org/wiki/PLY_(file_format)).
5. Burnard, L., and S. Bauman (eds.) *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. 2.6.0. Last updated on 20th January 2014. www.tei-c.org/P5/.
6. Ó Carragáin, É. (2007). *Christian Inculturation in Eighth-Century Northumbria: The Bewcastle and Ruthwell Crosses*. Colloquium Journal 4. ism.yale.edu/sites/default/files/files/Christian%20Inculturation%20in%20Eighth.pdf
7. Gabler, H. W. (2010). *Theorizing the Digital Scholarly Edition*. Literature Compass 7 (2). 43.
8. Siemens, R., et al. (2012). *Toward modeling the social edition: An approach to understanding the electronic scholarly edition in the context of new and emerging Media*. Literary and Linguistic Computing 27 (4). 445-61.

Simulating the Cultural Evolution of Literary Genres

Sack, Graham Alexander

gas2117@columbia.edu
Columbia University, United States of America

Wu, Daniel

danielwu@fas.harvard.edu
Harvard University, United States of America

Zusman, Benji

benji.zusman@gmail.com
University of Florida, United States of America

The evolution of literary form and style is an emerging area of academic research and offers a valuable case study in cultural evolution generally. Several notable papers have appeared recently. In particular, critic Franco Moretti has offered a number of provocative claims concerning the relationship between genre evolution and demographic changes in the 19th Century reading public:

- 1. Due to the growth of the reading public, the British novel underwent an abrupt change circa 1820: novels became far more heterogeneous and generically differentiated, aimed at specialized niches rather than readers in general.
- 2. The average lifespan of genres is ~25-30 years, the same as a human generation. This historical rhythm results from generational turnover in readers.
- 3. Literary genre evolution is characterized by alternating cycles of divergence and convergence—that is, periods of increasing generic diversity and differentiation followed by periods of decreasing diversity.

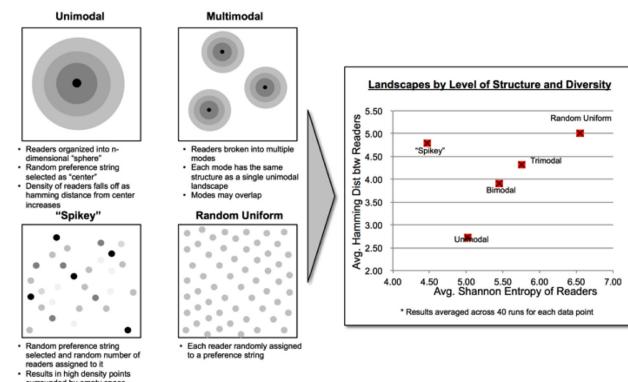
Statistician Cosma Shalizi argues in a response, "Graphs, Trees, Materialism, Fishing," that while Moretti identifies provocative historical patterns, he fails to fully articulate the mechanisms underlying and driving literary genre evolution.

The objective of this paper is to take up Shalizi's injunction by building a computational model of possible generative mechanisms driving genre evolution. We consider the following questions:

- How do the static characteristics and dynamic behavior of the 'reading public' affect literary genre evolution?
- How is generic diversity affected by reader diversity? Is there a *phase change* as the reading public grows?
- Under what circumstances will the life cycle of literary genres parallel the life cycle of generations?

We investigate these questions by constructing an agent-based model of two populations: (1) cultural forms (e.g., books); and (2) cultural consumers (e.g., readers). The key attribute of agents in each population is a bit string of user-specified length. For cultural forms, this bit-string represents the morphological features of the work: for instance, in the case of literature, bits represent attributes such as authorial style, length, plot, and theme.[1] For cultural consumers, the bit-string represents an individual's ideal preference. Each consumer has a tolerance for variation from this ideal represented as an acceptable hamming distance.

Individual cultural consumers are in turn organized into larger preference landscapes, which vary in their levels of structure, entropy, and reader diversity (see diagram).



Once the preference landscape has been constructed at set-up, a genetic algorithm is run on the cultural forms in order to simulate evolution. The fitness of each book is measured by the number of readers it receives in that time period.[1] High fitness books are more likely to survive and reproduce, increasing their influence on the content of the next generation of literary works. Three reproductive mechanisms are used:

- Survival: books carry over from generation T to T+1 with no change
- Asexual: individual bit-strings from generation T are copied with a user-specified probability of mutation to create a new generation of books at T+1
- Sexual: pairs of bit-strings from generation T are spliced in order to create a new generation of books at T+1

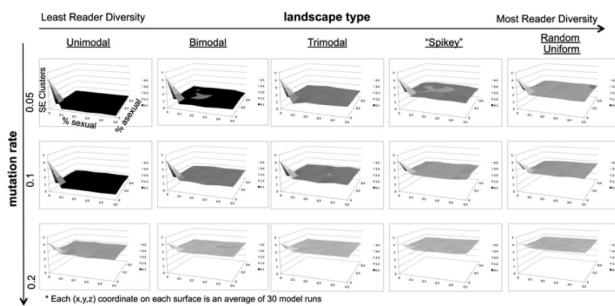
While the use of genetic and evolutionary paradigms to describe bibliographic change may at first seem suspect, each of these reproductive strategies has an intuitive interpretation in the context of literary production. Survival corresponds to the case in which market-successful books are simply reprinted. Asexual reproduction corresponds to the case in which successful books spawn similar works with slight variation: that is, authors copy and modify the template provided by recently successful works. Sexual reproduction corresponds to what we might call "genre-crossing": authors take the features of two successful works and synthesize them in order to produce a new work. The current trend of "mash-up" literature provides a salient example. Best-sellers such as *Abraham Lincoln: Vampire Hunter* splice the features of already-successful genres (e.g., historical biography and gothic). Lest we dismiss such works as gimmicks, it is worth recognizing that many high-prestige genres emerged through hybridization. Modernist works such as James Joyce's *Ulysses* self-consciously combined the features of the realist novel with those of the classical epic. *Pastiche*, *bricolage*, and the combination of high and low art were central to postmodern literature, epitomized by William Burrough's "cut-up" novels. Recombination is a widely-used mechanism in literary production.

The relative proportions of these reproductive strategies are parameterized variables, as is the mutation rate, which represents the probability that any feature of a work will be mutated during either reproduction process. The mutation rate also has an intuitive interpretation in the context of cultural production: it characterizes the average creative experimentalism of a particular cultural field, that is, how far authors are generally willing to depart from established models.

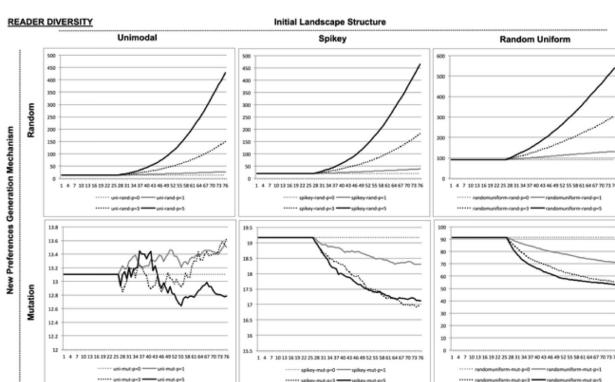
We run simulated experiments in order to determine the impact of various scenarios on literary genre evolution, including (i) variation in reader preference landscapes features, (ii) demographic changes such as population growth and generational turnover, and (iii) feedback between reader preferences and dominant cultural forms.

The results suggest a number of insights about plausible mechanisms driving the evolution of cultural forms generally and literary genre specifically.

First, generic diversity[1] cannot be explained solely in terms of the characteristics of the reading public: we also need to account for the characteristics of the creative process, in particular, the level of experimentation in the cultural market at a given historical moment, represented in this model by the mutation rate.



Second, contrary to Moretti's claim, we show that growth in the reading public is not sufficient to guarantee an increase in either reader diversity or generic diversity. In fact, market growth may actually reduce diversity under certain conditions. To determine the effect that growth will have, we need to know whether the preference landscape was initially homogeneous vs. diverse and whether new readers have preferences that are similar to or different from the readers who already populate that market.



Third, the model predicts that dramatic changes in the preferences of cultural consumers—analogous to ecosystem disruption—lead to increases in creative experimentation (i.e., the cultural mutation rate).

Lastly, we find that the preferences of conformist consumers have a highly disproportionate effect on the level of generic diversity relative to the rest of the consumer population, producing 'phase change' dynamics. Genres and cultural product categories tend to form around the preferences of conformist consumers because they have more reliable and predictable tastes.

Although the model above addresses a set of claims about literary genre, the implementation is intentionally general,

relying on abstract feature and preference strings that can represent any cultural product that can be atomized into variable features. Our intention in future research is to calibrate the model against case studies from a variety of cultural markets (literature, film, plastic arts, etc.).

References

- Daranyi, P. Wittek, L. Forro** (2012). *Toward Sequencing 'Narrative DNA.'* Proceedings Computational Models of Narrative. LREC. Istanbul.
- Hughes, J., Foti, N., Krakauer, D., Rockmore, D** (2012). *Quantitative patterns of stylistic influence in the evolution of literature.* PNAS, May 2012.
- Moretti, Franco** (2005). *Graphs, Maps, Trees.* Verso: New York.
- Rabkin, E. and Simon, C.** (2008) *Culture, Science Fiction, and Complex Adaptive Systems.* Biocomplexity at the Cutting Edge of Physics, Systems Biology, and Humanities. Bologna: Bononia University Press.
- Shalizi, Cosma** (2011). *Graphs, Trees, Materialism, Fishing.* Reading Graphs, Maps, Trees: Responses to Franco Moretti. South Carolina: Parlor Press.

Computational Models of Narrative: Using Artificial Intelligence to Operationalize Russian Formalist and French Structuralist Theories

Sack, Graham Alexander
Columbia University

Finlayson, Mark
MIT CSAIL

Gervas, Pablo
Universidad Complutense de Madrid

Panel Description:

Narrative has become an active research focus in the artificial intelligence community in recent years, with efforts aimed at the construction of computational models for storytelling and story understanding. AI researchers often speak of narrative as its own form of intelligence—a complex cognitive phenomenon encompassing natural language understanding and generation, common-sense reasoning, analogical reasoning, planning, physical perception, imagination, and social cognition. Computational models of narrative draw upon a multitude of different techniques from the AI toolkit, ranging from detailed symbolic knowledge representation to large-scale statistical analyses

Our purpose in proposing this panel for Digital Humanities 2014 is two-fold. First, we want to introduce literary critics and other humanities scholars to parallel efforts underway in AI-narrative research and computational modeling. Second, we want to open a dialogue about the process for and implications of operationalizing canonical narrative theories, especially those from Russian Formalism and French Structuralism, for the computational analysis and generation of stories.

We have selected the panel papers to represent two broad categories of research:

1. **Narrative Analysis:** use of computational techniques to identify, classify, or extract plot structures from a single text or a corpus of texts;
2. **Narrative Generation:** use of computational techniques to generate new stories based on story grammars, case-based reasoning, or other mechanisms

Computational narrative analysis dates at least as far back as the 1960s, when researchers such as Alan Dundes experimented with the automated classification of folktale themes using punch-card systems. Generative computational models date to the early 1970s, with systems such as TALE-SPIN. Most approaches to narrative generation have been based on either story-grammars or case-based reasoning: MINSTEL, BRUTUS, and MEXICA are canonical examples in the field. Recently, new approaches have leveraged crowd-sourcing and alternative generation mechanisms, such as games and network models.

We believe that this panel will be interesting and valuable for conference attendees both because of the innovative nature of recent AI-narrative research and because of its connection to the Russian Formalist and French Structuralist critical traditions. The papers by Finlayson and Gervás, respectively, adapt Propp's *Morphology of the Folktale* to the analysis and generation of narratives. Finlayson augments Propp with machine learning algorithms, while Gervás tests Propp's formalism with combinatorial simulations. Sack combines René Girard's classic theory of "triangular desire" in the 19th Century Novel with social network analysis to create a mechanism for the generation of narrative event sequences.

Operationalizing Formalist and Structuralist concepts by implementing them computationally both gives new life to canonical theories and raises a number of significant practical and theoretical questions. In what ways are Formalist story-grammars such as Propp's well-specified and in what ways are they under-specified? Can they be successfully generalized beyond their original domains? How can the analytical tools of literary critics be redeployed for generative purposes?

We hope that the panel will spur lively discussion around these and other questions.

Panel papers :

Paper #1: Learning Propp's *Morphology of the Folktale*

Abstract: Vladimir Propp's *Morphology of the Folktale* is a seminal and highly influential piece of work, and it is one of the most computationally-amenable theories of narrative structure that has been proposed. Until now we have not had the computational techniques that would allow us to learn Propp's theory automatically. I describe Analogical Story Merging (ASM), a machine-learning algorithm for extracting Proppian plot patterns from sets of stories. Remarkably, ASM can learn a substantive portion of Propp's theory of the structure of folktale plots. I will outline the abilities and deficiencies of the algorithm in detail. I will also discuss the data collection infrastructure that enables this work, namely, the Story Workbench, a general-purpose linguistic text annotation tool that supports the semi-automatic markup of over twenty different syntactic and semantic representations.

Bio: Dr. Mark Finlayson is a Research Scientist at the Computer Science and Artificial Intelligence Laboratory at MIT. His research focuses on representing, extracting, and using higher-order semantic patterns in natural language, especially focusing on narrative. He received the B.S.E from the University of Michigan in 1998, and the M.S. and Ph.D. from MIT in 2001 and 2011, respectively, all in Electrical Engineering and Computer Science. He is general chair of the Computational Models of Narrative Workshop series

Email: markaf@mit.edu

Homepage: www.mit.edu/~markaf

Paper #2: Generating Russian Folk Tales: A Computational Look at Some Aspects Propp Did Not Formalize

Abstract: Although it was never conceived as a computational framework, the semi-formal analysis of Russian folk tales carried out by Vladimir Propp has often been used

as theoretical background for the automated generation of stories. Many story generation systems attempted to generalize Propp's account to other types of stories or to combine it with other techniques. The added distance introduced by these extensions obscured the nature of the extensions required to transform an analytical view into a computational generative one. Deliberately constraining the domain to Russian folk tales, I revisit Propp's work to explore the gaps between Propp's original proposal and the requirements of a modern computational solution. These involve a number of procedures and considerations that Propp describes as fundamental to the process of story generation, but does not formalize. I will describe an existing system under development which respects the core concepts of Propp's morphology and extends them with additional computational elements for these missing aspects that are reverse engineered from Propp's descriptions and the examples in his book.

Bio: Dr. Pablo Gervás is an Associate Professor at the Facultad de Informática at Universidad Complutense de Madrid. He received his PhD from Imperial College in 1995. He has worked on natural language processing, computational creativity, and computational narratology. In the area of creative text generation, he has done work on automatically generating narrative, metaphors, and formal poetry. His current research focuses on studying the role of narrative in human communication, with a view to applying it in human-computer interaction. He is the director of the NIL research group (nil.fdi.ucm.es and also of the Instituto de Tecnología del Conocimiento (www.itc.ucm.es). Dr. Gervás has taken part in the organization of several scientific meetings on topics related to computational creativity, and he is currently involved in three projects funded by the European Commission on this topic, including the PROSECCO initiative (prosecco.computationalcreativity.net).

Email: pgervas@sip.ucm.es

Homepage: nil.fdi.ucm.es/index.php?q=node/92

Paper #3: Character Networks for Narrative Generation: Structural Balance Theory and the Emergence of Proto-Narratives

Abstract: This paper models narrative as a complex adaptive system in which the temporal sequence of events constituting a story emerges out of cascading local interactions between nodes in a social network. The approach is not intended as a general theory of narrative, but rather as a particular generative mechanism relevant to several academic communities: (1) literary critics and narrative theorists interested in new models for narrative analysis, (2) artificial intelligence researchers interested in new mechanisms for narrative generation, and (3) complex systems theorists interested in novel applications of agent-based modeling and network theory.

The paper is divided into two parts. The first part offers examples of research by literary critics on the relationship between social networks of fictional characters and the structure of long-form narratives. René Girard's theory of "triangular desire" in the 19th Century Novel serves as a key theoretical foundation. The second part provides an example of schematic story generation based on a simulation of the structural balance network model. I will argue that if literary critics can better understand sophisticated narratives by extracting networks from them, then narrative intelligence researchers can benefit by inverting the process, that is, by generating narratives from networks.

Author: Graham Sack is a Doctoral Candidate in English & Comparative Literature at Columbia University. His research focuses on the application of quantitative and computational methods to literary and cultural criticism, particularly the use of network analysis to study plot and characterization and the use of simulation to model the behavior of literary markets and the cultural evolution of literary genres. He holds an MA in English & Comparative Literature from Columbia University, an MSc in Economics from the London School of Economics, and a BA in Physics from Harvard College.

Email: gas2117@columbia.edu

Homepage: www.columbia.edu/~gas2117/grahamsack.html

Corresponding Panel Organizer:

Graham Sack
Doctoral Candidate
Columbia University
English & Comparative Literature Department
gas2117@columbia.edu
857-472-0062

References

- Dundes, Alan.** (1965). *On Computers and Folktales*. Western Folklore, Vol. 24, No. 3. 185-193

Meehan, J. R. (1977). *Tale-spin, an Interactive Program That Writes Stories*. In Proceedings of the Fifth International Joint Conference on Artificial Intelligence, 91-98. Los Altos, CA: William Kaufmann, Inc.

Turner, S.R. (1993). *Minstrel: A Computer Model of Creativity and Storytelling*. Ph.D. dissertation, University of California at Los Angeles, Los Angeles, CA.

Selmer Bringsjord and David A. Ferrucci. (1999). *Artificial Intelligence and Literary Creativity: Inside the Mind of BRUTUS, a Storytelling Machine*. Erlbaum.

Perez y Perez, R. (1999). *Mexica: A Computer Model of Creativity in Writing*. Ph.D. Dissertation, The University of Sussex, Falmer, UK.

Li, B., Lee-Ubran, S., Appling, D.S., Riedl, M.O. (2012). *Automatically Learning to Tell Stories about Social Situations from the Crowd*. Proceedings of Computational Models of Narrative, LREC 2012

Gervás, P. (2012). *Stories from Games: Content and Focalization Selection in Narrative Composition*. I Spanish Symposium on Entertainment Computing. Universidad Complutense de Madrid, Madrid, Spain

Propp, Vladimir (1968). *Morphology of the Folktale*. 2nd edition. University of Texas Press, Austin, TX.

Girard, René (1961). *Deceit, Desire, and the Novel: Self and Other in Literary Structure*. Johns Hopkins Press, Baltimore.

An Integrated Approach to the Procedural Modeling of Ancient Cities and Buildings

Saldana, Marie
marie.saldana@gmail.com
UCLA

I. Introduction

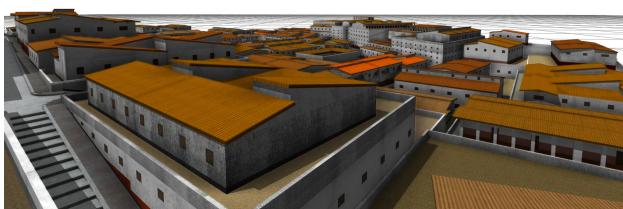


Fig.1: Urban infill created from the procedural 'domus' rule.

This paper presents a suite of procedural rules for creating 3D models of Roman and Hellenistic architecture and urban environments. The term 'rules' in procedural modeling refers to the computer code that generates a 3D model. Unlike traditional 3D modeling software such as SketchUp or 3ds Max, which use polygons to simulate form, procedural modeling entails the

use of computer programming languages in textual semantic description of a building that then generates a polygonal model. This represents not only a technical, but also an epistemological difference, as the choice of modeling method can influence not merely the cost or aesthetic outcome of a project, but also how information is selected, processed, and indeed what is considered to be information versus noise. Procedural modeling requires that each stage of the transmutation of data in the modeling process is rigorously thought out and documented, allowing 3D models to move beyond visualization to become robust research tools.

Procedural modeling has the potential to address a number of issues related to 3D archaeological reconstructions which are of concern to digital humanists. An important advantage of procedural methodology is that it allows for the rapid prototyping and interactive updating of 3D content. The use of attributes and parameters enables scholars to visualize change over time and gauge the impact of various factors on the built environment. Furthermore, these attributes and parameters can be tracked and harnessed as valuable geospatial data through the use of GIS software and of interactive visual displays. Of particular interest for archaeologists and architectural historians is the ability to test hypothetical reconstructions of ancient architecture in a fully realized urban context. Crucially for humanists, the procedural rules link each iteration of a model to its source material, allowing the degree of certainty present in each model to be accurately defined through the documentation of each step in the process of interpreting a given data set. Procedural modeling thus enhances the scholarly value of architectural reconstructions by providing a platform for the comparison and refutation of 3D visualizations.

II. Background and Software

Reconstruction of historical architecture and cities was an early application of procedural modeling. Significant test cases were built around ancient Rome and Pompeii.¹ However, the main commercially available procedural software behind these projects, ESRI CityEngine, is currently being marketed mostly as a low-cost rapid-prototyping pipeline to urban planners and production designers rather than to scholars for its ability to create data-rich, detailed architectural models.² Procedural engines are based on proprietary high-level graphics programming languages which are extremely time- and resource-consuming to produce for most academics, particularly in the humanities, and therefore few open-source alternatives exist.³ A precedent for my current project is Pascal Mueller's 2010 PhD dissertation which used classical temples as a case study for demonstrating the potential of CGA shape grammar, the procedural language that eventually became the core of ESRI CityEngine.⁴ I chose to use ESRI CityEngine for my work, because I find it to be the best commercial procedural modeling product for architecture and cities, and because its recent integration with ESRI GIS software such as ArcMap provides significant advantages for managing and visualizing archaeological data (Fig.2).

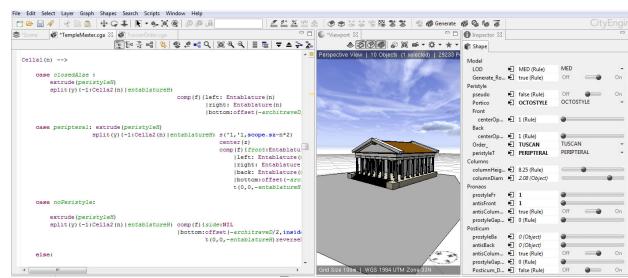


Figure 2: Screenshot from ESRI CityEngine showing text editor for rules (left) model (center) and model attributes (right).

III. Methodology and Research Application

My work aims at the creation of a full suite of procedural rules for the main typologies of classical architecture. A master

rule for classical temples exemplifies this project. The temple rule was designed to produce a schematic model of any kind of classical temple with a minimal number of parameters. By selecting from a few options, a user can instantly generate, for example, a tetrastyle Tuscan temple on a high podium; a peripteral Doric temple with pronaos and posticum; or an ionic pseudo-dipteral Hellenistic temple with variable intercolumniation. The rules were based on sources such as actual archaeological remains in Italy and Asia Minor, as well as Vitruvian templates. The rules are designed to be fully modular, that is, the rule for a specific typology such as a temple, arch, or stoa collates several sub-rules, which can be re-used and combined as necessary. These construct the components of a building, such as colonnades, entablatures, pediments, and roofs, or refer to a specific order, such as Tuscan, Doric, Ionic, or Corinthian. Urban models are then generated from geodatabases imported from GIS software ArcMap. These geodatabases contain the footprint of the building, along with specific attributes necessary to create the model (such as column diameter, order, and building type), and the bibliographic citations that reference the source material from which the attribute data was derived. All of this textual material may be queried in the final visualization of the model. The original impetus for the creation of the suite of rules was the generation of a series of 3D models of early Republican Rome, eventually to be visualized interactively with Unity game engine web viewer⁵. The procedural rules facilitated the hybridization of actual Roman data, comparanda from other sites, and hypothetical interpolation that was necessary to complete the picture of the Forum Romanum in this time period. Eventually, the project and the rules expanded to cover the site of Magnesia on the Meander in Turkey, necessitating the addition of new typologies. To date, the suite of procedural rules includes a core set of typologies responding to the topography of Rome and Magnesia on the Meander. These include temples, altars, basilicas, houses, shops, streets, triumphal arches, arcades, colonnaded streets, stoas, theaters, and stadiums (Figs.1,3,4). It is anticipated that these rules will become valuable tools for visualizing the urban fabric at numerous other locations where the archaeological data must be supplemented with well-researched templates that provide customizable parameters.



Figure 3: Procedural basilica, shown in a configuration with a row of shops in front

IV. Conclusions

Procedural modeling presents a powerful new methodology that has yet been underexploited by the Digital Humanities. Contrary to traditional 3D modeling methods, procedural modeling forces the investigator to approach visual and 3D content through a rigorously syntactic and process-oriented framework. This framework preserves the hierarchy of decisions that result in a given visual interpretation of archaeological evidence. Models thus produced are extremely information-rich and the ways in which they can be used to aid research are just beginning to be explored.

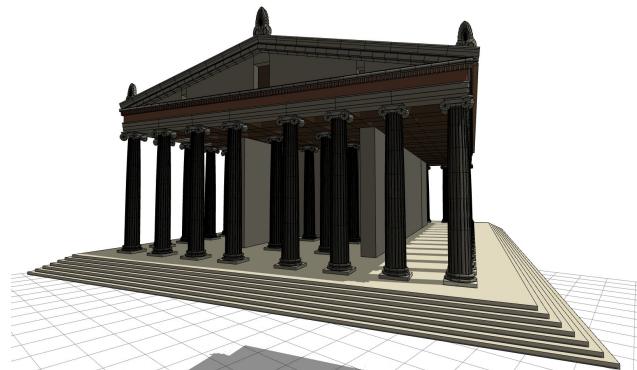


Figure 4: The Ionic temple of Artemis at Magnesia on the Meander, procedurally generated from the master temple rule.

References

1. See the Rome Reborn Project, romereborn.frischerconsulting.com/. The procedural aspects of this project were published in Dylla, Kimberly, Bernard Frischer et al., 2010. "Rome Reborn 2.0: A Case Study of Virtual City Reconstruction Using Procedural Modeling Techniques," in CAA 2009. Making History Interactive. 37th Proceedings of the CAA Conference March 22-26, 2009, Williamsburg, Virginia (Archaeopress: Oxford, 2010) 62-66. On Procedural Pompeii, see www.esri.com/software/cityengine/resources/casestudies/procedural-pompeii.
2. **ESRI CityEngine:** www.esri.com/software/cityengine
3. Some recent cultural heritage projects make use of parametric tools such as Building Information Modeling (BIM), for example ArchiCAD and Revit, or the Rhinoceros plug-in Grasshopper.
4. Mueller's (unpublished) dissertation and Parthenon rule, which is distributed as an example with CityEngine software, were indispensable in my efforts to master CGA shape grammar. However, his work was oriented to the field of computer science and restricted to peripteral temples of the Doric order. The rules I present here are entirely my own work, as a full restructuring and rewriting of the code, with the addition of much new material, was necessary to implement a wider agenda geared toward an architectural audience. For an overview of the architectural application of CGA shape grammar, see Mueller, et al., 2006. "Procedural Modeling of Buildings", in ACM SIGGRAPH 2006 Papers (ACM, Boston, 2006) 614-623.
5. This work was undertaken as part of the RomeLab project, an offshoot of the UCLA Experiential Technologies Center. The Unity models are playable at romelab.etc.ucla.edu/. RomeLab investigates the Roman Forum of ca. 186 BC, some 500 years before the period represented by Rome Reborn, therefore the procedural content of RomeLab is entirely new.

Digital Humanities Empowering through Arts and Music. Tunisian Representations of Europe through music and video clips

Salzbrunn, Monika

monika.salzbrunn@unil.ch
ISSRC-UNIL

Mastrangelo, Simon

simon.mastrangelo@unil.ch
ISSRC-UNIL

1. Introduction

How can digital humanities contribute to empowering processes through arts and music¹? In this long paper, we aim to analyse Tunisian representations of Europe through images, music and video clips. This in turn forms part of a broader research project funded by the Swiss National Science Foundation that investigates undocumented mobility in the context of recent developments in Tunisia. Even after the events of the "Arab Spring" and its demands for dignity and liberty, the desire of young men (*harragas*) to "burn their papers" (*harga*), to leave their country of first citizenship and to reach Europe still persists. However, such a desire to escape one's overall circumstances cannot be reduced to merely economic motives. Undocumented mobility is by no means a one-dimensional, single-layered process governed by "push-and-pull factors", but reflects the transnational social *imaginaire* and its various cultural resources as well.

1.1. Overview

New media such as social networks, blogs and YouTube, as well as their mobile symbolizations, sounds and images, contribute to the dissemination of mobilization, dissent and disagreement, creating a transnational² socio-cultural space and public spheres in which current (and past) situations are negotiated and contested³. Based on ethnographical fieldwork in Tunisia and Switzerland, as well as on digital anthropology of social networks⁴ and blogs⁵, this research project deals with the increasingly important role of these spaces and thus contributes to a fresh and innovative approach that relates undocumented mobility, (political) mobilization, transnational practices and the (gendered) social *imaginaire*.

1.2. Methodology

What do Tunisian migrants expect of the country they emigrate to? Do they hope to find a job easily, improving their living conditions as soon as they arrive in Europe? In this paper, we aim to reveal migrants' expectations of moving from Tunisia to Europe. We will take into account the representations of migrants before they leave their country of origin as well as those of migrants already living in Europe. We also seek to understand if these expectations are always the same or if they differ from one person to another. Can we identify certain mental representations that are widespread? By listing semiotic as well as audio-visual representations, we aim to understand what the most prevalent themes and images of migrations are.

Do mental representations evolve? When meeting people who recently emigrated to Europe, we will ask them if the way they now see Switzerland is the same as before. Are they satisfied with their new living conditions? Would they rather leave this country and give another country a try – Canada, for example? On the one hand, we will analyse the way people talk and write about these subjects on blogs, Facebook and YouTube. On the other hand, we will interview both Tunisians who have only been living in Switzerland for a few months and Tunisians who have been living here for much longer.

2. Getting Started

Our paper will take into account documented as well as undocumented migrations. In many videos posted on YouTube and Facebook, boat people explain why they emigrated. Sometimes the people we listen to are still waiting to leave, and sometimes they have already arrived in Lampedusa and are hoping to continue farther into mainland Europe. Finally, there are videos in which people who have arrived at their final destination tell us what they now think of undocumented mobility. Do they think they made the right choice? In this paper, we will analyse the vocabulary as well as the references they use in their discourse.

At the beginning of our research, we planned to analyse audio-visual material available 1) via YouTube and 2) via blogs. Fieldwork conducted in February 2014 has shown that it is more

crucial to focus on Facebook, since it seems to have a much wider audience than blogs on irregular migration.

The material analysed will be delimited by five key words in Arabic (different spellings will be taken into consideration) and French (with all possible combinations searched: k1; k1+k2; etc.). Firstly, an expert in media and information technology will help us perform computer-assisted content analysis. The second step is to add manual content and discourse analysis (of semiotics etc.) under the leading research question "How can digital humanities contribute to empowering processes through (popular) arts and music⁶ (namely Mizwoud⁷ and rap music⁸)?" We plan to conduct a general analysis of power relations later on, but within the limited space of the present paper, we will focus on processes of empowerment and general forms of representations of Europe.

Five kinds of video clips will be researched: a) music clips, b) music clips without video but with photos, c) news reporting, d) videos taken by harragas during their trips and e) other types of video. Our research schedule consists of four steps: 1) content analysis of the clips and comments posted by the author(s), 2) analysis of images, representations and symbols used in the videos, 3) analysis of the sound and the music, and 4) analysis of hypertext links and YouTube suggestions leading to other videos. This material will be juxtaposed to ethnography "in the real world": participant observation in the harraga milieu, semi-guided interviews and informal talks and life histories in Tunisia and Switzerland.

The desire for a better life abroad has been emphasized after the "Arab Spring" by the development of a transnational youth culture disseminated via Facebook, YouTube, and Twitter⁹. Representations of success in Europe also circulate thanks to transmigrants who bring back images and symbols of a higher social status to their home villages. These images and symbols are videotaped, transformed and used in music clips uploaded onto the Internet. New music styles have also emerged as part of a global vernacular language. An example is Mizwoud, which has partly become a language of the political resistance movement in the Maghreb countries. It also belongs to the "migrating population, and with them travelled outside Algeria into France, then Europe" (Nair 2007: 65). Hence, cultural resources¹⁰ have given rise to the development of a common artistic language of exclusion, dreams, representation of a better life and resistance¹¹. Videos, soundscapes and images bridge the gap between the street, the sea and the virtual, empowering highways and gates through which multiple belonging processes¹² take place.

References

1. Bally, John, Collier, Michael (2006). *Introduction: Music and Migration*, In: Journal of Ethnic and Migration Studies, vol. 32, no. 2, pp. 167-182.
2. Nair, Parvati (2007) *Voicing Risk: Migration, Transgression and Relocation in Spanish/Moroccan Rai*. In: I. Biddle/V. Knights (eds) *Music, National Identity and the Politics of Location: Between the Global and the Local*. Aldershot: Ashgate, 65-79.
3. Salzbrunn, Monika/Sekine, Yasumasa (2011) *From Community to Commonality. Multiple Belonging and Street Phenomena in the Era of Reflexive Modernization*. Tokyo: Seijo University Press.
4. Graziano, Teresa (2012): *The Tunisian diaspora: Between "digital riots" and Web activism*. In: Dana Diminescu (Hg.): *e-diaspora Atlas*. Paris: Maison des Sciences de l'Homme. e-diasporas.fr (27.2.14)
5. Héas, Stéphane, Poutrain, Véronique (2003) *Les méthodes d'enquête qualitative sur Internet*. In: ethnographiques.org/2003/Heas_Poutrain.html (27.2.14)
6. Bally, John, Collier, Michael (2006). *Introduction: Music and Migration*, In: Journal of Ethnic and Migration Studies, vol. 32, no. 2, pp. 167-182.

7. **Stapley, Kathryn** (2006): *Mizwid: An Urban Music with Social Roots*, in: Journal of Ethnic and Migration Studies, vol. 32, no. 2, pp. 243-256.
8. **Friese, Heidrun** (2012) 'Ya l'babour, ya mon amour' – *Rai-Rap und undokumentierte Mobilität*. In: Marc Dietrich/Martin Seeliger (Hg.) Deutscher Gangsta-Rap. Sozial- und kulturwissenschaftliche Beiträge zu einem Pop-Phänomen. Bielefeld: transcript, 231-85.
9. **Najar, Sihem** (2011) *Mouvements sociaux en ligne, cyber activisme et nouvelles formes d'expressions en Méditerranée*. Institut de recherche sur le Maghreb contemporain, Bulletin 6:3.
10. **Salzbrunn, Monika** (2011) *Mobilisation des ressources culturelles et participation politique: l'apport des cultural studies à l'analyse des rapports sociaux dans un contexte festif*. Migrations et Société (édité par Daniel Bertaux/Catherine Delcroix/Roland Pfefferkorn), 23, 133: 175-92.
11. **Souiah, Farida** (2011): *Musique populaire et imaginaire migratoire en Algérie*. In: Diversité, no. 164, pp. 27-33.
12. **Salzbrunn, Monika/Sekine, Yasumasa** (2011) *From Community to Commonality. Multiple Belonging and Street Phenomena in the Era of Reflexive Modernization*. Tokyo: Seijo University Press.

Digital humanities in Estonia: digital divide or linguistic isolation?

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

Kulasalu, Kaisa
 kaisa.kulasalu@gmail.com
 Estonian Literary Museum

1. Introduction

The paper introduces the disciplinary developments in Estonian humanities that are intertwined with use of digital resources and methods. Estonia forms an interesting case. On the one hand, the post-Socialist country is well-known as an example of technological innovation, it was the first to introduce paperless governmental processes, the electronic ID-card is compulsory document and its electronic functions are widely used, to name a few examples.¹ However, at the same time, the developments in the field of digital humanities have not taken place at the same pace with the ones in USA or Western Europe.

The aim of this paper is to compare and contrast the use of digital technologies in Estonian humanities to the developments of the field of digital humanities.

2. Use of the digital technologies in Estonia

In 2012, 78% of the population of Estonia aged 16-74 years used the internet. There are various electronic services, like e-Tax Board where 93% of income tax declarations have been made in 2012. From 2005, it is possible to participate in elections electronically. In October 2013, 21,2% votes in the local elections were given electronically. And what is more, 62% of the inhabitants took part of the 2011 Population and Housing Census electronically.² However, the use of digital solutions is not a spotless success story: according to PIAAC (Programme for the International Assessment of Adult Competencies) study in 2011-2012, 30% of 16–65 year old test group refused to take a part of the test online and 13% did not succeed in the simplest computer-related tasks. The results of PIAAC study reveal a digital divide that is related to the linguistic divide: there are no significant differences in gender, education or the social background, but Estonian-speaking group had in general better skills in the use of technological tools than the group of the speakers of other languages.³

Apparently, Estonian society is not able to provide all its IT-advantages in the languages other than Estonian enough to integrate the non-Estonians to the Estonian IT-world, thus leaving them isolated to a certain extent. Despite being the wired country in terms of the e-Government, not all fields of life have been equally keeping up the developments in the IT, the humanities being one of these areas. The National Strategy for the Development of Information Society until 2020 addresses the problem of inequality and of moderate specific IT skills in the fields other than ICT.⁴

3. Field of digital humanities in Estonia

Schnapp and Presner⁵ have distinguished between two waves of digital humanities: quantitative and qualitative one. According to the divide, the first wave was about digitisation projects and creating infrastructure, the other wave consists of interpretation and research methodologies for digitised and born-digital materials. Taking this approach as a framework, Estonia could be described as being in the first wave of digital humanities. In archives and libraries, there have been large-scale digitization projects from early 2000s onwards, the digital archival systems and repositories have been created and made publicly available for many different collections.

For example, in Estonian Literary Museum, a file repository and archival information system Kivike⁶ has been created to manage the collections to store the digital materials and metadata, and to make the collections as much as possible publicly available via Internet. New archival data is being added to the online repository and described there, among other materials the born-digital sources (photographs, audio and video recordings, as well as "digital manuscripts") that have been collected systematically since 1994.

This illustrates the tendency of including some of the notions of second wave of digital humanities, because the emphasis on preserving of born-digital materials has also been apparent in several institutions. It is partially connected to the practical needs of document management in a country where governmental practices have moved online. The digital preservation department in Estonian National Archives was founded in 1999 with the goal of "permanent preservation of digital data despite the changes in society and technology". Web pages are also digitally archived: Estonian National Library is in charge of creating and maintaining a web archive of the web resources that are important for the Estonian culture. For born-digital materials, these approaches still focus on creating an infrastructure rather than using the data. Creating and maintaining the collections have been in the centre of digital humanities projects in Estonia, using them for educational or research purposes is scarcer.

One of the most clearly developed disciplines that uses digital data sets and methodology is linguistics. There has been a constantly developing co-operation of the linguists and computer-scientists since the 1950s in the field of language technology, there has been formed a specific curriculum of computational linguistics at the University of Tartu in the 1997⁷, large corpora and dictionaries are in constant progress and in 2006 the special national financing program has been founded in order to promote the development of language technology and digital language resources.⁸

Except of the linguistic studies, the use of digital methods in the humanities research has been somewhat sporadic and depends of the interests and skills of individual researchers. In the curriculums of humanities there are no comprehensive courses on digital research methodology (the social sciences with its traditional data analysis are in much more better situation here). In some of the humanities fields digital methods are being taught in specialised courses. For instance, students of archival science need to follow the course on digital preservation.

To sum up the situation in Estonia, the digital infrastructures for the major humanities collections in Estonia have been created by now or are in progress. There are several original web solutions for presenting and/or collecting the humanities data, a couple of successful examples of crowdsourcing have

taken place. The creation of the software for the audience in humanities has been more than modest. Because of the linguistic restrictions, and the small size of researchers' community the specific software (e.g. for the analyse of the Estonian folk song melodies) would be tailor-made for the needs of (the group of) individual researchers and is not developed for the public (or multilingual) use. Use of the digital methods in the research has been constantly increasing, but only occasionally, the lack of systematic technological knowledge base seems to be crucial. All this does not apply for linguistics. However, the research mostly takes place within the disciplinary boundaries and the collaboration between the digital humanists of different disciplines has been modest, without signs for the need of the umbrella discipline.

Only in October 2013, first seminar of digital humanities in Estonia took place providing an overview of the projects in progress within the field.⁹ Different fields of humanities were represented: linguistics, archeologists, literary studies, folkloristics, arts. In the presentations, digital humanities in general got little attention, most of the presenters focusing on digitizing, maintaining and/or presenting the collections (either digitized or born-digital) rather than interpreting the data. In addition to this a couple of research projects were presented, as well as an technologically innovative film project, two software projects for documenting the artefacts, and a few initiatives of using open data in web projects. It is clear, however, that the seminar did not involve all the digital humanities projects in course. For example there was no paper from Estonian National Archives, although it is serving as the competence center in the field of the digital preservation in Estonia, and there are several IT projects ongoing.

After the seminar an informal network was created with a mailing list and homepage. According to a small survey taken in early 2014, the humanities scholars who work with digital technologies feel the need for collegiality in the field, i.e. for wider digital competence among their colleagues both for developing new tools and using the already existing ones.

4. Discussion

Digital humanities became a discipline because the humanities scholars as users of digital resources needed to understand the digital mechanisms and get some academic recognition for using them. One of the crucial questions of the field is how to bring the technological advantages and knowledge within the reach of researchers so they could develop the tools and environments they need. To code or not to code, that is the question, when being probably the only person in the world who would need a language-specific tool, or if the audience of a web environment would consist of mere ten people. The size of the linguistic community restricts the reasonable amount of work hours spent for the IT solutions or web-pages. On the other hand, the available English-language software is often either not known or not easily applicable for several reasons. This is how the linguistic divide becomes the digital divide even in the country with well-established technological infrastructures. The awareness and the use of various IT solutions in Estonian society with its linguistic divide between Estonians and non-Estonians follows the model of the situation in the digital humanities leaving the peripheral group to the isolation to some extent due to the language barrier.¹⁰

References

1. Herlihy, Pete (2013). 'Government as a data model': what I learned in Estonia, accessed 1. November 2013. digital.cabinetoffice.gov.uk/2013/10/31/government-as-a-data-model-what-i-learned-in-estonia/
2. Eesti infoühiskonna arengukava 2020, accessed 7.3.14. infoyhiskond.eesti.ee/files/Info%C3%BCChiskonna%20arengukava%202020.pdf
3. Täiskasvanute oskused Eestis ja maailmas. PIAAC uuringu esmased tulemused, accessed 1.11.2013. www.hm.ee/index.php?popup=download&id=12426
4. Eesti infoühiskonna arengukava 2020, accessed 7.3.14. infoyhiskond.eesti.ee/files/Info%C3%BCChiskonna%20arengukava%202020.pdf
5. Schnapp, J. and Presner, P. (2009). 'Digital Humanities Manifesto 2.0', accessed 31. October 2013. www.humanitiesblast.com/manifesto/Manifesto_V2.pdf.
6. Kivike, accessed 1.11.13. kivike.kirmus.ee
7. Koit, Mare (1998); Õim, Haldur. Arvutuslingvistika mujal ja meil. In: Keel ja Kirjandus, No 1, pp. 1-7.
8. Eesti Keeletehnoloogia. National Programme for Estonian Language Technology, accessed 1.11.13. www.keeletehnoloogia.ee
9. Kulasalu, K., Sarv, M. (eds.) 2014. *Digitaalhumanitaaria Eestis Ao 2013 - Digital Humanities in Estonia Ao 2013*, www.folklore.ee/dh/DHE2013/
10. Galina, Isabel (2013). *Is There Anybody Out There? Building a global Digital Humanities community*. Accessed 1.11.13, humanidadesdigitales.net/blog/2013/07/19/is-there-anybody-out-there-building-a-global-digital-humanities-community/

"How To Do (Digital) History" and Undergraduate Digital Humanities

Schell, Justin

sche115@umn.edu
University of Minnesota

Gabaccia, Donna

drg@umn.edu
University of Minnesota

Where exactly is the place of digital humanities to be in undergraduate education? If, indeed, 21st century universities must begin to prepare students for professional work in which digital familiarity, skills, and facility are increasingly central, where is the site of responsibility for that training? It could be disciplinary or interdisciplinary, located within the curricula and pedagogies of existing departments or relocated to the information sciences, library, or core liberal education curriculum.

Joining a chorus of scholars considering the place of undergraduate pedagogy and digital humanities, including those on DH2013's "The Future of Undergraduate Digital Humanities" panel, this presentation will detail the highly collaborative creation and facilitation of "How To Do History," a course offered by Donna Gabaccia. In its old form, the course was one of the mainstays of upper-level offerings by the University of Minnesota's History department, serving as a way to prepare students' Senior Thesis and, implicitly, prepare these students for graduate school in History; it has at least temporarily become a venue for students explorations of, and contributions to, digital history in a disciplinary context unofficially known as "How To Do (Digital) History."

The core collaboration is between the presenters, respectively, a well-established Professor of History (Gabaccia) and the Digital Humanities Specialist for the University of Minnesota Libraries (Schell). In the summer of 2013, we met to discuss not just the content of the class (readings, assignments, etc), but also what kinds of support the Library could make available to students to actually make projects (rather than, as in the past, preparing to write individual research papers). By circumscribing the options available to them (limiting them to a few of the main digital humanities tools and methods, i.e. mapping, Omeka, digital storytelling), it made both the technologies, and the projects that could be created with those technologies, much more accessible. In addition to creating public projects, the students engaged critically with reframed historiographical questions (i.e. the writing and rewriting of history through Wikipedia) and digital literacy (critically examining Twitter and other online platforms of communication and how they relate to scholarly discourse), reading, among others, Cohen (2006)¹, Rosenzweig (2011)², Kelly (2013)³, and Nawrotszi and Dougherty (2013)Nawrotszi, K. and

Dougherty, J. (2013). *Writing History in the Digital Age*. Ann Arbor: University of Michigan Press.⁴

While collaboration was essential to the undergraduates' creation of digital projects, graduate students also became partners in the course. Instead of Gabaccia trying to supervise the seven groups of undergraduate students, she allowed graduate students to enroll in their own section of the course. In addition to doing further readings with her regarding the intellectual lineage and stakes of digital history and digital humanities (e.g. Gold 2012)⁵, the graduate students served as project managers for the undergraduate groups, getting valuable experience in facilitating collaborative projects and various digital humanities tools and methods, neither of which are normally part of their History graduate work.

A further collaborative aspect to this version of "How To Do History" was extensive collaboration with the University of Minnesota Libraries. While many subject librarians will come once during the semester to introduce research methods, subject-specific database, and other source materials, Schell attended nearly every class, doing multiple presentations about specific ideas and technologies, ranging from introductions to significant digital history projects over the last 15 years to demonstrations of specific tools. In addition, he met multiple times individually with both graduate students (to help construct a blogging assignment for the undergraduate students) as well as the collaborative project groups, helping to refine the scope their projects to make them manageable as a single semester of work.

Finally, one project group worked specifically with the Upper Midwest Jewish Archives, part of the Libraries' Archives and Special Collections, creating an Omeka exhibition around previously unprocessed and undigitized materials. The group looked at the lives of two Jewish men in the early 20th century, through two World Wars, work in the printing industry, and global travel, all set against the backdrop of vicious anti-Semitism in Minneapolis, characterized at that time as the most anti-Semitic city in the United States. A second group created a web feature for the University's James Ford Bell Library about changing cartographical representations of Scandinavia by mapmakers in the premodern world.

The lessons from "How To Do History" do not end with the completion of these collaborative, public digital history projects. Reflecting on the course after its completion, we realized that, due to the project- and group-based environment, many students considered "middle of the road" in their skills and interests developed more research and technological skills than in previous iterations of the class taught by Gabaccia. While most students are not opting for a digital Senior Thesis, instructors have relayed to us anecdotally that they see a greater preparedness and skill in terms of research in those who took the digital version of "How to Do History" than other versions of the course. Furthermore, as we noted above, the graduate students who supervised and facilitated the undergraduate projects gained valuable experience that opened new directions for their own graduate work in terms of research and instruction. Reactions at the departmental level have been mixed. The History Department proudly featured the student work on its webpage but also continued its ongoing internal debate about the future of HIST 3959—is the course necessary? Could methodologies other than digital history be featured? Senior members of the department also continued to express skepticism about their ability to evaluate digital work; the department has responded by offering a future department-wide workshop on that issue.

The knowledge gained from organizing and teaching the course can inform the development of similar courses in different departments (e.g., How To Do Digital Sociology, Anthropology, Ethnomusicology, to name a few). Schell's position in the Libraries facilitates that possibility. Digital literacy is a critical element to any undergraduate education, regardless of discipline, and especially if one seeks to receive graduate training in that field. Integrating these digital humanities lessons within each discipline helps students engage more deeply in the development of critical inquiry and with the specific transformations underway in these fields. Furthermore, it creates opportunities for transdisciplinary education, as the digital tools and methodologies students used to create these

projects are not just limited to digital history projects. Whether this be part of an emerging "digital humanities" cohort or a broader idea of "digital studies," it allows for collaborative relationships where extra-disciplinary institutions, such as Libraries, are essential partners.

References

1. Cohen, D. (2006). *Digital History: A Guide to Gathering, Preserving, and Presenting the Past on the Web*. Philadelphia: University of Pennsylvania Press.
2. Rosenzweig, R. (2011). *Clio Wired: The Future of the Past in the Digital Age*. New York: Columbia University Press.
3. Kelly, T., (2013). *Teaching History in the Digital Age*. Ann Arbor: University of Michigan Press.
4. Nawrotzki, K. and Dougherty, J. (2013). *Writing History in the Digital Age*. Ann Arbor: University of Michigan Press.
5. Gold, M., ed. (2012). *Debates in Digital Humanities*. Minneapolis: University of Minnesota Press.

Intellectual Property Rights vs. Freedom of Research: Tripping stones in international IPR law

Scholger, Walter

walter.scholger@uni-graz.at

Center for Information Modeling - Austrian Centre for Digital Humanities, University of Graz

The paper will address some of the most common and frequent needs and obstacles regarding legal issues in current digital scholarship (e.g. ownership of digital copies, electronic provision of source material) and demonstrate some of the consequent misconceptions, restrictions and legal traps which result from the lack of legal certainty due to the heterogeneous international legal situation regarding IPR and ancillary copyright.

While the free availability of sources has been a long-lasting demand and desire in all fields of research, open access to the results of scientific research has become a de facto obligation in recent years. This is reflected in the requirements of many national and international funding bodies demanding the public and free availability of research results and publications.

Generally speaking, humanities research focusses on products of the human mind – hence, the research object is usually subject to intellectual property rights. Largely based at universities, cultural heritage institutions or other public research institutions, that research is usually non-commercial and based on a public mandate for education, with little to no funding available for the acquisition of licenses and the proper remuneration of IPR holders. Open (and free) access to sources – especially those available only in cultural heritage institutions like archives and libraries – gains further importance because national funding agencies (e.g. the Austrian FWF) generally do not allow for the inclusion of license fees in their grants.

On the other hand, researchers themselves have a keen interest in defending their own intellectual property rights, in part due to economic concerns but also in terms of academic credit. This conflict of interest is evident even in the Universal Declaration on Human Rights, Art. 27, which sets the premise that "(everyone has) the right freely [...] to share in scientific advancement and its benefits", but goes on to say that "everyone has the right to protection of the moral and material interests resulting from any scientific, literary or artistic production of which he is the author".

Several European Council Directives (2001/29/EC, 2003/98/EC, 2004/48/EC) have made a strong case for public access and free use of educational and scientific resources. This political agenda has been visible in the 7th EU framework and is also evident in several UNESCO publications (e.g.

"Recommendation concerning the Promotion and Use of Multilingualism and Universal Access to Cyberspace" and the "Charter on the Preservation of Digital Heritage", both dating back to 2003). However, the national implementation of these ideals is lagging behind: The actual legal situation regarding the use of and access to digital resources in many member states of the European Union (and the UNESCO, respectively) poses a number of difficulties.

While Common Law legal systems – most prominently the US and the UK with their allowances for Fair Use and Fair Dealing – focus more on society's interest in the access to and use of publications for education and self-improvement, the Civil Law systems found in continental European countries stress the rights of authors. Therefore, the usage, distribution and especially the electronic provision of resources require distinct free licenses, i.e. privileges for the educational sector. For non-digital material, a tried and trusted system of such privileges has been in place for decades. However, many countries – Austria among them – have so far failed to implement the necessary legal changes to extend these licenses to digital sources.

This ambiguity between the treatment of non-digital and digital resources poses another problem: Most humanists (or scholars in general, regardless of their respective domains) are unfamiliar with the legal implications of their work. Often drawing assumptions based on long-standing experiences and practices with non-digital material, few are familiar with the details of current legislature on digital sources. Also, though notable and admirable exceptions exist, there is generally also little to no support from universities' legal offices.

Where source material is owned by universities or cultural heritage institutions, or has moved into the public domain due to the expiration of applicable protection periods (usually 70 years for printed materials), humanities scholars have little need to address such concerns. But more recent sources – especially when dealing with the current interest in Big Data – pose a number of legal challenges. Also, orphaned works (works without a known and retraceable author), while at first glance not subjected to the usual IPR restrictions, are dealt with very differently in the various European countries, but usually involve collecting societies.

Furthermore, cultural heritage institutions often insist that the ownership of physical resources automatically induces a right to their digital copies and demand fees based on that claim – which may, in fact, be unfounded, either because the digitization (which in itself is not subject to IPR but rather of ancillary copyright) was not done by the institution or other legal obligations (e.g. in national or regional archival laws) oblige them to freely provide material in the public domain. The increase of collaborative work across not only disciplinary but also national borders adds another dimension to the already puzzling situation: Which legal system applies to resources hosted in different countries and which legal framework must be used for electronic publications?

A number of these questions can be addressed by taking a look at international IPR treaties like the Berne Convention for the Protection of Literary and Artistic Works, the Trade Related Aspects of Intellectual Property Rights (TRIPS) or the World Intellectual Property Organization (WIPO) Copyright Treaty.

The paper will therefore demonstrate some of the most common legal obstacles that humanities scholars encounter in the course of digital research and teaching and try to provide an overview of the current legal situation, the differences and common denominators of Civil Law and Common Law systems regarding IPR (especially the electronic provision of material) and the international framework of Copyright treaties. Since an exhaustive juxtaposition of international legal differences is unfeasible due to scale, the Austrian example will be used to showcase some of the most obvious and momentous shortcomings of current IPR legislature, and will be compared with German law (which has addressed some of these issues while still keeping strong restrictions in place) and the Anglo-American concepts of Fair Use and Fair Dealing.

In conclusion, the paper will try to define a possible best practice which draws on the analysis of the common denominators found in international treaties, UK and US law

and the EC Digital Agenda for Europe as expressed in recent EU directives.

References

- Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonization of certain aspects of copyright and related rights in the information society.
- Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the re-use of public sector information (amended 2013).
- Directive 2004/48/EC of the European Parliament and of the Council of 29 April 2004 on the enforcement of intellectual property rights.
- Goldstein, P.** (2010). *International Copyright. Principles, Law, and Practice*. New York: Oxford University Press.
- Jahnel, D.** (2012). *IT-Recht*. Vienna: Verlag Österreich.
- Kuhlen, R.** (2008). *Erfolgreiches Scheitern – eine Götterdämmerung des Urheberrechts?* Bozenburg: VWH.
- Schöwerling, H.** (2007). *E-Learning und Urheberrecht an Universitäten in Österreich und Deutschland*. Vienna-Munich: Verlag Medien und Recht.
- Torremans, P.** (2007). *Copyright Law. A Handbook of Contemporary Research*. Cheltenham: Edward Elgar Publishing Ltd.

Revisionism as Outreach: The Letters of 1916 Project

Schreibman, Susan

Trinity College Dublin

The Letters of 1916 Project is the first crowd-sourced digital humanities project in Ireland. The project pivots around one of the most important events in twentieth-century Irish history -- the 1916 Easter Rising -- in which a small group of Irish Volunteers rebelled against the British Army on 24 April, Easter Monday 1916. The Rising was quickly quashed by the British, but the executions of its leaders several weeks later made martyrs of them, and set in motion a series of events that resulted in Irish independence from the United Kingdom in 1922.

The Letters of 1916 project focuses on this tumultuous period. Its goal is to create a crowd-sourced digital collection of letters written for a six-month period before and after the Easter Rising (1 November 1915 – 31 October 1916).

The project includes letters held at institutions (in Ireland and abroad), alongside those in private collections. The collection criteria is extremely broad: it includes any letter written to or from someone in Ireland, or that contains Irish subject matter.

The ultimate goal is to create a new online resource that will add a novel and heretofore underutilised viewpoint to the events of the period, a confidential and intimate perspective that will form the basis for a revision of our understanding of life in Ireland in the early 20th Century.

Letters 1916 integrates four distinct theoretical and methodological approaches:

- The creation of a thematic research collection
- A crowd-sourced public humanities project
- Text analysis
- Social media for dissemination and outreach

This paper will explore phase I of the project: the sourcing, gathering, and transcribing/encoding of collection. It will also document how the project is being used in the classroom at the Masters level.

Phase II will transform the transcribed/encoded letters, along with their accompanying images, into a thematic research collection where users will be able to do full text searching, text analysis, view correspondences, view letters temporally and spatially, and engage with thematic exhibitions. The collection is being viewed as a major new resource and as such, will

ultimately be deposited at a cultural heritage institution for long-term preservation.

In her 2009 article Rose Holley made the distinction between social engagement projects, which she defined as ‘giving the public the ability to communicate with us and each other’ via such activities as tagging, commenting, rating, and reviewing, and crowdsourcing which uses social engagement techniques to help a group of people achieve a shared, usually significant, and large goal by working collaboratively together as a group. Crowdsourcing also usually entails a greater level of effort, time and intellectual input from an individual than just socially engaging.

The project stops short, however, of a collective editing project. Phase 1 of the project was launched in September 2013 to a great deal of press.[1] The database was pre-populated with some 300 letters from cultural heritage institutions. To date, the project has 1500 letters in its workflow (about 800 publicly available).

The project, thus far, uses two distinct crowdsourcing methodologies according to the typology developed by Carletti, et al: a) correction and transcription; and b) complementing collection. Building on the immense success of the Europeana World War 1 Road shows, we invited the public to bring their letters in for scanning at the launch. This model is being replicated, thus far in Ireland and the UK. Moreover, the software allows the public to upload their letters from home, create basic metadata, and provide the project with descriptive information.

At the time of this writing, there are approximately 30 privately-held collections (or letters from small institutions we would have never have thought to reach out to) for letters from this period. The size of individual deposits range from one letter to 90– a correspondence between the donor’s grandmother and grandfather who happened to be courting during 1916.

The public is also invited to transcribe previously uploaded letters. The response has been so overwhelming that the project team is frequently can barely keep up with demand. The engagement, encouragement, and reaching out to volunteers is a major aspect of the project to be discussed. At present the site does not utilise some of the more traditional features of crowdsourced projects (such as badges and icons) to engage its volunteers. However, other methods have been implemented, such as a volunteer forum, featuring contributors both on the project site and in the more traditional forms of media.

The project has benefited immensely from previous crowdsourced projects, from its methodological approaches, to its workflow and software. The project utilises WordPress for its front end, and for its content management system a version of Omeka that was modified by University of Iowa libraries (DIY History) to support crowdsourcing projects. It utilises George Mason’s Scripto tool for transcribing, and the Transcribing Bentham’s TEI/XML toolbar for light encoding. In a domain space that has constantly called for the creation of tools, the ability to fairly rapidly and without a great deal of bespoke programming to string together these diverse tools into one unified web presence, may signal a golden age of tool development for our field.

The Transcribing Bentham toolbar has been an interesting feature which many transcribers have utilised with little instruction. The Transcribing Bentham project designed the toolbar to hide ‘much of the complexity of TEI markup from the transcriber’ (Moyle, 353). An analysis of the markup used in this environment reveals an especially intuitive use of tagging, something the TEI is rarely praised for. Despite the caveat that the toolbar only allows the encoding of 15 out of the TEI’s hundreds of tags, the ease in which individuals with absolutely no experience in textual editing, AML, or text encoding have been utilizing it may point to new ways to create a more intuitive encoding interface. A downside to the reduced set of tags available is that users, in trying to make sense of the tags available to them, use certain tags, such as <lb> (line break), so ubiquitously that it almost amounts to tag abuse.

This project is pitting, in many respects, methods typical of thematic research collections— including the creation of a carefully curated dataset with fully transcribed and proofed documents, and hand-crafted markup and metadata created

in accordance with the Text Encoding Initiative Guidelines-- against the less painstaking creation of a full text dataset amenable to text analysis and visualisation as a method of discovery and analysis. Ultimately, in Phase II of the project, we intend to offer readers both views – the traditional reading environment of the thematic research collection along with the analytics available via text analysis and visualisations. But the decisions we make at this stage which will be detailed in this talk will impact what is possible in later phases.

The social media dimension of finding and engaging our audience demonstrates just how significantly the environment in which digital humanists operate has changed over the past decade. An analysis of the first month’s tweets reveal an ever expanding network of cultural institutions, individuals, organisations, and traditional media who have passed the message of the project along. A twitter-feed is a feature of the project’s home page and discussions amongst the project team reveal a shifting notion of how to position the project in a web of similar initiatives in Ireland coming broadly under the rubric of The Decade of Commemoration.[2] The project has already been positioned not as a silo, but as an aggregator embracing multiple communities of holders of these precious objects. We share back with individuals and institutions photographs we take, and will eventually share the fully transcribed/encoded texts files. But as we near the centenary of the Easter Rising, we are aware of our responsibility to, on the one hand, maintain the integrity of the project while, on the other, opening our resource more fully to sister initiatives, while respecting copyright and intellectual property promises to our donators.

References

- Carletti, Laura, Gabriella Giannachi, Dominic Price, Derek McAuley,** (2012). ‘*Digital Humanities and Crowdsourcing*’. MW2012: Museums and the Web 2013. Online
Causer, Tim, Justin Tonra, Valerie Wallace (2012). ‘*Transcription Maximized; Expense Minimized? Crowdsourcing and Editing The Collected Works of Jeremy Bentham*’. Literary and Linguistic Computing. 27/2 2012. 119-137.
Dunn, Stuart and Mark Hedges. ‘*Crowd-Sourcing Scoping Study: Engaging the Crowd with Humanities Research*’. Centre for e-Research, Department of Digital Humanities, King’s College London. Online.
Holley, Rose (2010). ‘*Crowdsourcing: How and Why Should Libraries Do It*’. D-Lib Magazine. March-April 2010. 16:3/4. Online
Moyle, Martin, Justin Tonra, Valerie Wallace (2011). ‘*Manuscript Transcription by Crowdsourcing; Transcribing Bentham*’. Liber Quarterly 20:3/4. March 2011.
Palmer, Carol (2004). ‘*Thematic Research Collections*’. Companion to Digital Humanities’. Blackwell. Online.
Robinson, Peter. (2010). ‘*Editing Without Walls*’. Literature Compass. Special issue on Scholarly Editing in the Twenty-first Century. 57-61.

Please see letters1916.ie for full press coverage
This shorthand refers to a decade of significant historical events in Ireland 1913-1923 in Ireland, beginning with the Dublin Lockout in 1913 and ending with Irish independence, followed by a Civil War a decade later.

Digitizing the Dead and Dismembered: DH Technologies for the Study of Coptic Texts

Schroeder, Caroline T.

University of the Pacific, United States of America

Zeldes, Amir

Humboldt University, Berlin

The Coptic language evolved from the language of the hieroglyphs of the pharaonic era and represents the last phase of the Egyptian language. It is pivotal for a wide range of humanistic disciplines, such as linguistics, biblical studies, the history of Christianity, Egyptology, and ancient history.

Whereas languages like Classical Greek and Latin have enjoyed advances made in digital humanities with fully-fledged online research environments accessible to students and scholars (such as the Perseus Digital Library), until recently, no computational tools for Coptic have existed. Nor has an open digital research corpus been available. The research team developing Coptic SCRIPTORIUM (**S**ahidic **C**orpus **R**esearch: **I**nternet **P**latform for **I**nterdisciplinary **m**ultilayer **M**ethods) is developing and providing open-source technologies and methodologies for interdisciplinary research across multiple disciplines in the Coptic language. This paper will address the automated tools we are developing for annotating and conducting research on a Coptic digital corpus.

Conducting digitally-assisted and computational research in Coptic using available DH resources is complex for several reasons. Most texts are preserved from damaged, incomplete, and dismembered manuscripts or papyri. The DH project papyri.info has begun to create an online open-access resource for the study of Greek papyri and is beginning to digitize Coptic papyri and ostraca (ancient pot-shards with writing). These texts, however, are primarily documentary, consisting of wills, contracts, personal letters, etc. Coptic literary and monastic texts, the core of Coptic SCRIPTORIUM, are essential for the study of the Bible, intellectual history, literary history, and religious history. The manuscripts containing these texts were removed from Egypt in the seventeenth through nineteenth centuries piece by piece (sometimes page by page). Some have been published, many have not, and very few have been digitized in a format suitable for digital and computational work. Texts must be reconstructed from pieces of manuscripts published in fragments and/or stored in various libraries and museums worldwide. The status of Coptic literary and monastic complicates metadata management and corpus architecture: what constitutes a "work" – the codex in which a copy of the text appeared (and which may be dispersed across multiple physical repositories)? the manuscript fragment housed in a particular library or museum repository or the work, which only might survive in fragments of multiple codices (all copies of a "book" from the monastery's library), and thus in fragments not only from more than one codex but also more than one modern repository?

Coptic scholarship still lacks many standards for digital publication and language research that are taken for granted in Greek and Latin. As with other ancient languages, Coptic manuscripts are written without spaces. However, in contrast to its ancient counterparts, scholarly conventions on word division differ substantially from scholar to scholar. Additionally, since Coptic is an agglutinative language, the relevant unit for linguistic analysis is the morpheme, below the 'word' level. This means that segmentation guidelines must be developed for both levels of resolution. In order to search multiple texts, guidelines and tools for normalization, part-of-speech tagging and lemmatization of Coptic must be developed. These tools need to take into account Coptic's agglutinative nature, e.g. normalizing and annotating on the morpheme and word levels.

Finally, the development of the Coptic language during Egypt's Greco-Roman era raises questions about the origins of the language, its usage in a multilingual context, and the language practices of its ancient speakers and writers.

Coptic consists of Egyptian grammar, vocabulary, and syntax written primarily in the Greek alphabet; some Egyptian letters were retained, and some Greek and Latin vocabulary was incorporated into the language. The richness of the vocabulary's languages of origin varies from author to author, genre to genre. And despite recent publications on the topic, much research remains to be conducted on the extent and nature of multilingualism in late antique Egypt, especially during the fourth and fifth centuries. Additionally, due to the agglutinative nature of the language, one word can be comprised of morphemes with different languages of origin.

This paper will focus on the automated tools our project is developing to process the language, especially tokenizing

and part-of-speech annotations. Coptic SCRIPTORIUM has developed the first tokenizer and part-of-speech tagger for the language, and in fact for any language in the Egyptian language family. The presentation will address the unique challenges to processing and annotating the Coptic language.

We will present our current technical solutions, their accuracy rates, and the potential for future research. We will also address the ways in which this language's and corpus's unique features differentiate them from other more widely studied ancient languages, such as Greek and Latin. Examples will be drawn from the open-access corpora we are developing and annotating with these tools, available at coptic.pacific.edu (backup site www.carrieschroeder.com/scrptorium). The Coptic corpora processed and annotated with these tools can be searched and visualized in ANNIS, a tool for multi-layer annotated corpora. We anticipate this presentation to be of interest to scholars in digital humanities working with ancient languages and manuscript corpora as well as DH linguists and corpus linguists.

References

- Bentley Layton** (2011), *A Coptic Grammar*, 3rd Edition, Rev, Porta Linguarum Orientalium Neue Serie 20 (Wiesbaden: Harrassowitz, 2011), 19–20.
Layton, *Coptic Grammar*, 5.
J. N. Adams, Mark Janse, and Simon Swain (2002), *Bilingualism in Ancient Society* (Oxford: Oxford University Press, 2002); Arietta Papaconstantinou, ed., *The Multilingual Experience in Egypt from the Ptolemies to the Abassids* (Burlington: Ashgate, 2010).
<http://www.sfb632.uni-potsdam.de/annis/>

Progress Through Regression. Modeling Style across Genre in French Classical Theater

Schöch, Christof

christof.schoech@uni-wuerzburg.de
University of Würzburg, Germany

Riddell, Allen

allen.riddell@dartmouth.edu
Dartmouth College, USA

1. Introduction

Considerable scholarship in stylometry has focused on authorship attribution. Such work is based on the assumption that rates of high frequency "function" words (in contrast to "content" words) are reliable clues to authorship and are largely independent of factors like theme or genre¹. More recently, focus seems to have moved beyond the most frequent words to involve all vocabulary appearing in a corpus^(2, 3, 4). As many of these words vary strongly by context, factors like theme, genre, literary period or literary form have received greater attention.

This paper makes two contributions. First, we test the hypothesis that authorial style depends on genre and find that this is indeed the case, even when only considering the most frequent words. Second, in light of this result, we argue that adding additional features such as genre to a familiar model of authorship attribution offers a useful and novel way to investigate how authors' writing varies depending on context. We demonstrate how stylistic analysis making use of more articulate probabilistic models might move beyond established but limited models such as principal component analysis and distance-based clustering and achieve a better fit between model and hypothesis.

2. Data

In French literary studies, there is longstanding interest in analyzing the formal and stylistic constraints associated with classical theater^(5, 6). Playwrights from this period, such as Pierre Corneille and Jean Racine, figure prominently in early quantitative work in French literary studies, predating the use of digital computers^(7, 8). Whereas this pioneering research focused on single texts or a single author's works, today's availability of a wide range of digital texts, of flexible tools, and of vastly increased computing power permits more complex methods of analysis.

We have chosen to work on a corpus of 108 plays in three genres written by eight authors. The plays were produced over a period of roughly five decades (1630-1678) and the authors were selected because they wrote several plays in more than one genre. Table 1 illustrates the distribution of the plays across authors and genres.

	comedy	tragi-comedy	tragedy
Corneille, Pierre	9	1	20
Corneille, Thomas	8	0	15
Du Ryer	1	7	6
Molière	7	1	0
Quinault	1	1	3
Racine	1	0	9
Rotrou	1	4	3
Scarron	8	2	0
Totals	36	16	56

All texts are taken from the "théâtre classique" collection⁽⁹⁾ and have been preprocessed to include only character speeches.¹⁰ In order to better explore the variability of writing found among the authors and genres in the corpus, each play has been split into approximately 1,000 word sections. After processing, the corpus used for analysis contains 1,605 sections. Only the most frequent 100 function words in the corpus are retained.¹¹

3. Hypothesis and Method

Our hypothesis is that authorial style varies depending on genre. In order to test this hypothesis, we compare three models that predict the author of a section based on word frequencies and the genre of the section. The first model predicts the author based on word frequencies alone, ignoring information on genre. The second model adds to the first rudimentary information about how likely authors are to appear in each genre. The third model differs from the second in that it predicts the author of a section based on word frequencies for each genre separately. If authorial style varies by genre, then the third model should perform significantly better than the first model.

All three models are multinomial logistic regressions.¹² Multinomial logistic regression has been used for authorship attribution before¹³, but our approach expands on this by examining the use of a non-traditional covariate such as genre. Our aim is to encourage the building of interpretable models in order to understand how variables such as genre influence authorial style.

In statistical terms, the first model includes a global intercept parameter and word frequencies as predictors. The second model adds a genre-specific intercept parameter. The third model differs from the second model by allowing the regression coefficients associated with word frequencies to be different depending on the genre. These models may be expressed symbolically as shown in Fig. 1 (where the text section is

indexed by i and $\text{softmax}_k(a)$ is the extension of the inverse logistic function to multiple categories).

$$\text{Model 1: } \Pr(y_i = \text{author}_k) = \text{softmax}_k(\alpha + x_i\beta)$$

$$\text{Model 2: } \Pr(y_i = \text{author}_k) = \text{softmax}_k(\alpha_{\text{genre}[i]} + x_i\beta)$$

$$\text{Model 3: } \Pr(y_i = \text{author}_k) = \text{softmax}_k(\alpha_{\text{genre}[i]} + x_i\beta_{\text{genre}[i]})$$

$$\text{softmax}_k(a) = \frac{\exp(a_k)}{\sum_{j=1}^K \exp(a_j)}$$

Fig. 1: Three models of logistic regression

A point estimate for the parameters is obtained by maximizing the likelihood function using numerical methods. Models are fitted using randomly selected sections corresponding to four-fifths of the corpus. Models are then compared by measuring their out-of-sample predictions: an error rate for each model is calculated on the remaining fifth of the play sections, asking the model to predict the sections' authors based on word frequencies and, where applicable, genre information. This procedure is repeated fifty times, each time randomly partitioning the corpus.¹⁴

4. Results

The error rates associated with each model are shown in Fig. 2.

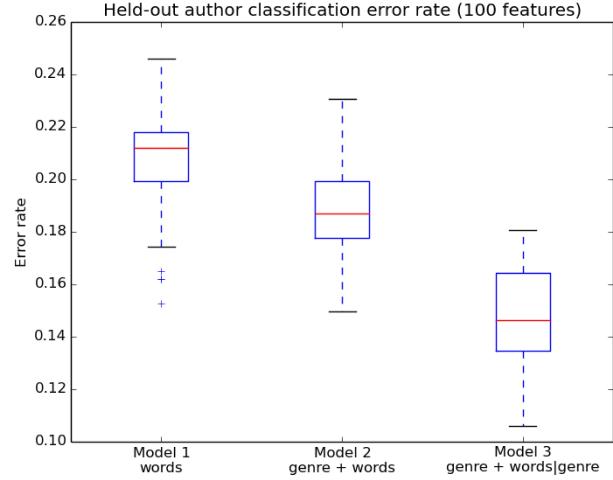


Fig. 2: Held-out author classification error rate (100 features)

In 49 out of the 50 trials, model 3 had the lowest error rate. In this corpus and for these authors, there is therefore little doubt that authorial style varies by genre. Table 2 shows the average error rates by model and genre.

	Model 1	Model 2	Model 3
Comedy	0.24	0.23	0.19
Tragi-comedy	0.25	0.22	0.10
Tragedy	0.17	0.15	0.13

5. Discussion

The variation of authorial style by genre underlying these results is best illustrated by looking at the frequencies of selected words that depend on both author and genre. For example, a few words are used with consistency across genres by one author but in another author vary considerably depending on genre. Table 3 indicates relative frequencies for three such cases.

	Pierre Corneille: comedy	Pierre Corneille: tragedy	Thomas Corneille: comedy	Thomas Corneille: tragedy
"est"	22.0	20.9	31.6	24.7
"par"	6.4	6.4	6.3	9.4
"au"	5.1	5.9	6.1	6.3

The auxiliary "est" and the preposition "par" are both used consistently across genres by Pierre Corneille but with a widely varying frequency between comedy and tragedy by Thomas Corneille, while the opposite behavior is true of "au". The preposition "par" is associated very frequently, in Thomas Corneilles plays, with causality (reason or effect) linked to emotions or moral principles (par bonté, par la gloire, par le respect). While the auxiliary "est" (third person singular present tense of "être") has an even more elusive semantic charge, it is mostly associated, in Thomas Corneille's plays, with statements of fact. Both phenomena seem to indicate a greater reliance, by Thomas Corneille, on causal relations and factuality in the tragedies than in the comedies, whereas the same contrasting treatment cannot be observed in Pierre Corneille.

The existence of such variation points to two notable facts. First, and contrary to common understanding, some very frequent function words other than personal pronouns do vary with genre within the work of a given author. Second, whether this is the case does not depend on the word in itself, but may differ from author to author. Therefore, such words are not exclusively or inherently markers of genre. Even when using only the very most frequent function words and even when excluding personal pronouns, then, authorship attribution cannot rule out that some influence from genre also comes into play.

On a different level, an explanation for the better performance of model 3 over model 1 brings in contextual information from literary history. Tragedies are usually described as being more closely bound to conventions of the "doctrine classique" than comedies or trag-comedies (¹⁵, ¹⁶). Therefore, the range of vocabulary and the pattern of usage would be expected to be more predictable in tragedy than in other genres. Were this indeed the case, a model might achieve a lower variance in its predictions by considering tragedy separately. This hypothesis is difficult to evaluate as it is difficult to "hold constant" authorship; authors tend not to write in equal amounts in different genres.

A critical explanation of model 3's superior predictive performance would point out that the task of predicting an author on the basis of word frequencies might change dramatically depending on the authors being compared. It might therefore be suggested that the better performance obtained by model 3 reflects this fact more than it reflects within-author variation across genre. In response to this criticism, it should be observed that model 3 performs better even when the same authors are being compared; Pierre Corneille and Thomas Corneille dominate numerically the samples from comedies and tragedies. Furthermore, the words shown in table 3 demonstrate that there is variation within an author's style across genres. Model 3 is designed to use this variation to attribute authorship.

6. Conclusion

We offer the following conclusions from this experiment. First, authorial style does appear to vary with genre even when considering only the 100 most frequent words. This suggests that factors such as genre should be systematically taken into account for authorship attribution. Second, logistic regression is a useful method in this context and should be part of the stylometric toolbox as it permits a range of information to be modeled jointly with authorship. Logistic regression could also be used to test for further relevant factors beyond genre, such as form (e.g. verse and prose) or theme (e.g. historical plays vs. religious plays).

References

- Hoover, D.** (2004). *Testing Burrows's Delta*. LLC 19. 453-475.
- Hoover, D.** (2004). *Testing Burrows's Delta*. LLC 19. 453-475.
- Burrows, J.** (2007). *All the Way Through: Testing for Authorship in Different Frequency Strata*, LLC, 22. 27-47.
- Rybicki, J. and Eder, M.** (2011). *Deeper Delta Across Genres and Languages: Do We Really Need the Most Frequent Words?* LLC 26: 315-21.
- Bray, R.** (1927). *La formation de la doctrine classique en France*. Paris.
- Scherer, J.** (1951). *La Dramaturgie classique en France*. Paris: Nizet.
- Muller, Ch.** (1967). *Étude de statistique lexicale: le vocabulaire du théâtre de Pierre Corneille*. Paris: Larousse.
- Bernet, Ch.** (1983). *Le Vocabulaire des tragédies de Racine*. Analyse statistique. Geneva / Paris: Slatkine / Champion.
- Fière, P.**, ed. (2007-2013). *Théâtre classique*, www.theatre-classique.fr.
- Speaker names, stage directions, dramatis personae, prefaces, metadata and other paratextual elements have been excluded from the analysis. Trailing sections having fewer than 500 words were discarded. Trailing sections having between 500 and 1,000 words were normalized and put in terms of rates per 1,000 words.
- Because of the relatively small sample size, three content words that may have an association with a specific genre ("coeur", "amour", and "yeux") appeared in the initial list of the top 100 most frequent words. These words were removed from the vocabulary so that the corpus contained only function words. The 100 most frequent graphical words used are: a, ai, au, autre, aux, avec, bien, c, ce, ces, cet, cette, comme, d, dans, de, des, donc, dont, du, elle, en, enfin, est, et, faire, fait, faut, grand, ici, il, j, jamais, je, l, la, le, les, lui, m, ma, mais, me, mes, moi, moins, mon, même, n, ne, non, nous, on, ont, ou, où, par, pas, peu, peut, plus, point, pour, puis, qu, quand, que, quel, quelque, qui, quoi, rien, s, sa, sans, se, ses, si, son, sont, suis, sur, t, tant, toujours, tous, tout, trop, tu, un, une, veux, voir, vois, vos, votre, vous, y, à, être.
- Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B.** (2003) *Bayesian Data Analysis*. 2nd ed. Chapman and Hall/CRC, 2003, pp. 430-33.
- Madigan, D., Genkin, D., Lewis, D.D., and Fradkin, D.** (2005). *Bayesian multinomial logistic regression for author identification*. Bayesian Inference and Maximum Entropy Methods in Science and Engineering, 803, 509-516.
- It is worth noting the complexity of the models considered here. Model 3 has 2424 parameters (each genre has an intercept for each author and an 8 by 100 matrix of author-word coefficients). Fitting the model requires maximizing a function with 2424 parameters, something that was challenging a decade ago. To be clear, the maximization is not taxing; it requires roughly 300M of memory.
- Bray, R.** (1927). *La formation de la doctrine classique en France*. Paris.
- Scherer, J.** (1951). *La Dramaturgie classique en France*. Paris: Nizet.

Hartmut Skerbisch – Envisioning association processes of a conceptual artist

Semlak, Martina

martina.semlak@uni-graz.at

Centre for Information Modelling, University of Graz

The paper will present the digital genetic edition of the notebooks of the Austrian conceptual artist Hartmut Skerbisch (1945-2009). The edition is currently emerging as part of a PhD project which explores the use and advantages of applying semantic technologies to art historical source materials.

Therefore, the central research question is how a digital genetic and semantically enriched edition can support the appreciation of an artist's concepts and associations during the process of a work's creation. Such insights should facilitate the interpretation and comprehension of his work.

Hartmut Skerbisch dealt conceptually with an extended notion of sculpture and space and applied these ideas to his works. He explored the relation between these entities, especially how space in general is generated or changed through the interaction of sculpture with audience and setting (e.g. how can electricity generate or compress space) and how electronic media change our perception of space.^{1 2} In his notes, he mentioned different authors, philosophers and novelists like James Joyce, Franz Kafka, Kathy Acker or Rudolf Steiner, as well as musicians and bands, ranging from classical to rock to blues music. He frequently cited passages from their works, commented on their ideas and theories, and consequently reflected them in his artistic expressions.

The 35 notebooks originated between 1969 and 2008 and have a scope of around 2100 pages. They include handwritten texts, calculations, formulas and sketches. The major editorial challenge regarding this source material is dealing with painted over or torn out passages, deletions and corrections. Skerbisch's notes do not follow a linear construction, they are often fragmentary and lacking formal notation. Additionally, the interconnection between text, graphics, annotations and references to other text passages itself generates an artistic composition in the two-dimensional medium of the notebook. At first glance, texts and sketches seem to be randomly arranged on the surface: Closer inspection, however, reveals an underlying principle of the different building blocks, whose form, shape, size and placement give the entire text additional meaning. This feature of the notebooks calls for special attention to a document-oriented view in addition to a work-oriented view.³

Consequently, the paper will discuss three key issues:

1. The need for editions of (hand)written sources in art historical research is evident. The recent editions of the sketchbooks by Max Beckmann⁴ as well as the class notes on form and design theory by Paul Klee⁵ show that traditional philological edition methods are not entirely sufficient for art historical needs. Issues of text-image relationships - especially for these types of sources, where images are of equal value as texts or text passages themselves become graphic elements - have not been adequately resolved.
2. Writing a notebook is a process manifested in time and space, therefore the evolution of a text over time is an important aspect. A genetic edition focuses on the document aspect and attempts to trace the origination process with a view on corrections, deletions and additions of the text to envision the author's original intentions. Especially modern manuscripts, with fragmentary and cursory notes, often exist as unfinished drafts and put different demands on the editors.^{6 7 8} The difficulties with digitally encoding such handwritten sources of visual artists with TEI are addressed: The paper will report on the benefits and drawbacks of the already implemented elements and attributes for genetic editions in the TEI guidelines and demonstrate their application on the material at hand.
3. Finally, the paper focuses on how semantic technologies could help to uncover the cultural and intellectual background of Hartmut Skerbisch's creative work, through relating textual references to concepts. Recent research projects like the Theodor Fontane's notebooks⁹ pursue a similar approach and emphasize the importance of linking concepts for semantic analysis of the source material.

Thus, in addition to the content-related and formal structure of the text, recurring concepts like people, places, works of literature, music and film are annotated. For uniquely referencing such entities, controlled vocabularies as well as authority files such as VIAF, GND and GeoNames are used. Beyond that, the aforementioned concepts are further classified. The identified concepts are extracted from the TEI document, mapped to a RDF model and stored in a triple store to facilitate reasoning.¹⁰ In this manner, implicit relationships between these

concepts are established which are not explicitly present in the texts. For example, a quotation from a novel is linked to that novel's author and other works of the same author. Thus, it is possible to compare and juxtapose the concepts used which allows for a variety of different analyses of the content.

The advantages of these methods are a) bringing the fragmentary entries of the notebook into a content-related sequence, b) visualizing unexpected combinations and associative chains of concepts to show the many-faceted influences on the artist's work, c) establishing a relation between the notebook entries and Skerbisch's realized works of art and finally d) tracing the creative process to promote the understanding of his work. The insights gained in the process will provide a basis for further research questions and analysis in art history and digital edition.

References

1. **Fiedler, E.** (2005). *Hartmut Skerbisch*. Sphäre 315. www.museum-joanneum.at/de/skulpturenpark/skulpturen/hartmut-skerbisch-1?overlay=true (accessed 7 March 2014).
2. **Fenz, W.** (1994). *Hartmut Skerbisch*. Werkauswahl 1969-1994. Graz: Neue Galerie.
3. **Robinson, P.** (2013). *Towards a Theory of Digital Editions*, Variants – Journal of the European Society for Textual Scholarship, 10: 105-131.
4. **Zeiller, C.** (2010). *Max Beckmann. Die Skizzenbücher*. Ein kritischer Katalog. München: Hatje Cantz.
5. **Zentrum Paul Klee** (2011). *Paul Klee – Bildnerische Form- und Gestaltungslehre*. www.kleegestaltungslehre.zpk.org (accessed 7 March 2014).
6. **Brüning, G., Henzel, K. and Pravida, D.** (2013). *Multiple Encoding in Genetic Editions: The Case of 'Faust'*, Journal of the Text Encoding Initiative. jtei.revues.org/697 (accessed 7 March 2014).
7. **Burnard, L., Jannidis F., Pierazzo, E. and Rehbein, M.** (2008-2013), *An Encoding Model for Genetic Editions*. www.tei-c.org/Activities/Council/Working/tcw19.html (accessed 7 March 2014).
8. **Pierazzo, E.** (2009). *Digital Genetic Editions*. The Encoding of Time in Manuscript Transcription. In: Deegan, M. and Sutherland, K. (eds.), Text Editing, Print and the Digital World. Farnham: Ashgate, 169-186.
9. **de la Iglesia, M. and Göbel, M.** (2013). *From entity description to semantic analysis: The case of Theodor Fontane's notebooks*. In: Ciotti, F. and Ciula, A. (eds.), The Linked TEI: Text Encoding in the Web. TEI Conference and Members Meeting 2013, Rome: Universitalia, 24-29.
10. **Allemand, D. and Hendl, J.** (2011). *Semantic Web for the Working Ontologist*. Effective Modeling in RDFS and OWL. Burlington: Morgan Kaufmann.

The potential of open computer-mediated communication channels to facilitate collaboration in geographically distributed collaborations

Siemens, Lynne
siemensl@uvic.ca
 University of Victoria

1. Introduction

1.1. Overview

Digital Humanities (DH) is becoming an increasingly global community of practice¹ with international initiatives such as centerNet², Global Outlook::Digital Humanities³, the Alliance of Digital Humanities Organizations (nd), the Digging into Data Challenge (2013) and many others. With advances in telecommunications and information technology, these types of collaborations are no longer bound by geography. However, as documented elsewhere, challenges stemming from geographical distance must be managed to ensure that teams are work together successfully. One of the primary challenges is finding ways to facilitate communication and coordination across distance and time (Olson & Olson, 2000; Siemens, 2010b; Siemens & Burr, 2013). Skype and other internet-enabled tools provide some potential to accomplish this; however, little knowledge exists on the best way to use these tools within a geographically dispersed collaboration. This paper will contribute to this discussion with an examination of the experiences of a DH lab using an open communication channel through Skype to connect team members located in different sites.

As Olson and Olson (2000) outline, despite advances in information and telecommunications technology, distance between members still impacts on a team's functioning at the task and personal levels. As they suggest and confirmed by others (Kennedy, Vozdolska, & McComb, 2010; Kraut, Galegher, & Egido, 1987), some amount of social presence or visible awareness of others is needed to allow for the sharing of advice, feedback and support among each other and the teams as a whole. Team members build on this social presence to task coordination. For co-located teams, face-to-face formal meetings and informal interactions in common rooms and around the proverbial coffee pot is the primary way to create and reinforce social presence. (Belanger & Allport, 2008; Kennedy et al., 2010; Warkentin & Beranek, 1999). For dispersed teams, the challenge to the creating this awareness exists because fewer channels and fewer personal cues in the communication exist. Without these, individuals are less likely to pay attention to each other (Short as cited in Warkentin & Beranek, 1999). Less personal forms of communication such as email tends to produce an "out of sight, out of mind" effect. (Hiltz as cited in Warkentin & Beranek, 1999). So, the question is how might computer-mediated communication like an open audio and video communication channel, through something like Skype, overcome distance and sustain social presence so that the team can achieve its tasks?

1.2. Methodology

This project grew from the desire of a DH lab that wanted to connect members who were split between two locations. As one initiative, the lab wanted to experiment with an open audio and video communication channel between the two offices. Using Skype, cameras and monitors would be installed and operating during work hours, allowing for communication between the two sites.

In order to understand the effectiveness of this communication channel in facilitating task and personal relationships, lab members were interviewed on two occasions.

The interview questions focused on the participants' understanding and experiences of this type of technology to facilitate collaboration between geographically disbursed sites (Marshall & Rossman, 1999; McCracken, 1988). The first round occurred before the cameras and monitors were installed (pre interviews). The second happened after the communication channel had been in place for several weeks (post interviews).

1.3 Findings

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.

As found in the pre interviews, these participants had little to no experience with this type of communication channel and were not sure what to expect. At the same time, they could see its potential for increased collaboration because they would be able to see their colleagues who located elsewhere and could more instantly communicate, much like if the person was seated next to them. While some participants expressed some concerns about privacy, as a whole, they were more curious about the ways in which the communication channel would actually work. Some of the questions focused on the location of the monitors and cameras, hours of operation, camera sight lines, and the ways in which communication would be facilitated. Overall, they were intrigued and excited to see how this open communication would work and support the lab's work.

After the cameras and monitors had been in place for several weeks, the lab members relayed that they became quickly accustomed to the presence of the cameras and monitors. In fact, when asked, they had difficulty recalling the day that when these were installed and the communication channel was opened. As several noted, the cameras and monitors were perceived to be "just there". In terms of challenges, noise was an issue from the outset. The microphones amplified all sounds that caused much distraction. As a result, the microphones were turned off after several days. Participants also noted that uneven coverage of team members due to camera placement. Some were very visible while others sat outside the sightlines.

Despite this, the participants were very positive about the experience and the benefits produced. Many noted that this open communication channel reinforced the feeling of collaboration by providing an "extension of the existing space" and a "hole in the wall" to the other members. As a result, the lab felt less divided by distance. The interviewees noted several benefits. First, because they could constantly see each other, members were reminded of the presence of those in the other office. In some cases, those in the other office might come join conversations that they had seen on the monitor. Second, an opportunity was created to model professional and academic work habits, such as reading, writing, thinking and discussing, which reinforced these for the others. And while the lab is a professional space, the participants also had a sense of play with each other. They would wave at each other and make visual jokes.

1.4 Conclusion

While this paper reports on the experiences on a small DH lab, it suggests potential for other geographically dispersed collaborations. Open audio and video communication channels can create the sense of social presence by reminding members that they are part of larger efforts, even working at a distance. These tools also complements the other well-established online ones, such as basecamp, github, email, and others, as well as face-to-face meetings for project coordination and decision-making (Ruecker, Radzikowska, & Sinclair, 2008; Siemens, 2010a; Siemens & Burr, 2013; Siemens, Cunningham, Duff, & Warwick, 2011).

References

1. Siemens, L., & 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.
2. centerNet. (nd). About Retrieved October 6, 2011, from digitalhumanities.org/centernet/about/
3. Global Outlook::Digital Humanities. (2013). Global Outlook: Digital Humanities, October 28, 2013, from www.globaloutlookdh.org

Pelagios 3: Towards the semi-automatic annotation of toponyms in early geospatial documents

Simon, Rainer

AIT Austrian Institute of Technology

Barker, Elton T. E.

The Open University

de Soto, Pau

University of Southampton, United Kingdom

Isaksen, Leif

University of Southampton, United Kingdom

Introduction

Since place names form the underlying semantic content of almost all geographic documents, the ability to identify them in texts and images is essential in any attempt to work with, compare or interpret them. For early maps and geographic texts this ability is especially important, because while they rarely conform to standard geometries or schemas, they often provide the earliest attestations to towns, peoples, and other spatially localized phenomena. Tools, infrastructure and resources for collating, aligning, and exploiting toponyms in early maps and geographic documents would therefore have a broad and significant impact across a range of fields, including Archaeology, History, Classics, Genealogy and Modern Languages.

In this paper we showcase early work on the detection of possible toponyms in digitized texts and scanned old maps. It builds upon the successful Pelagios initiative which has been connecting a variety of heterogeneous online resources related to classical antiquity. In contrast, Pelagios 3 will extend its scope to the European, Islamic and Chinese Middle Ages, but focus predominantly on geographic works. These in turn will form a core body of material around which we hope to see a more diverse body of references accumulate in time.

Since our ultimate aim is to enable humanities scholars to annotate and discover places in documents for themselves, we discuss our use and adaption of existing open source tools within a framework that puts a premium on flexible, lightweight and easy to use resources. Moreover, that discussion will be based on two real-case scenarios, in order to demonstrate the strengths of our approach and flag up potential issues that require further attention.

Mapping places from texts: the Vicarello Goblets

Our first test case tackles the issue of extracting place names from a text. The Vicarello Goblets are a collection of four silver drinking vessels dated to around late third or early fourth century AD, engraved a land itinerary between Gades (modern Cadiz, Spain) and Rome. Each goblet indicates the road stations along the route (varying between 104 and 110 on each goblet), as well as the distances in miles between them. These unusual ‘texts’, the limited range of places that they represent, the easy identification of the majority of locations, along with the fact that there are images and transcriptions available online already, makes the Vicarello Goblets an optimal source for trialling the methodology that we will use on much larger corpora of texts, including travel guides, gazetteers, encyclopaedias and more.

To obtain annotations from the Vicarello Goblets, all the toponyms are matched against places in a URI-based gazetteer, that is, a directory of places which assigns a persistent Web address to each entity, allowing for disambiguation at a global level. The engravings on the Vicarello Goblets already represent ordered lists of toponyms. Therefore we can directly match the lists against the gazetteer

based on name similarity, and then disambiguate by taking into account geographic proximity between different places in the list. For further documents which are of a less structured nature (i.e. which contain more free-form narrative text) we are experimenting with a combination of ‘geoparsing’ technologies, including the Edinburgh Geoparser and the Stanford NLP Toolkit.

Identification is only half the story, however. The data model that we have developed for Pelagios 3 allows for rich item metadata that cleanly differentiates between information about the item and information about the places that relate to it (and how). For instance, toponyms in a document may follow a certain sequence or layout. A simple mashup not only shows the toponyms from the four Vicarello Goblets on a map, but how they relate to one another as an itinerary. An information box at the bottom provides the information about the document itself (Fig. 1), while a small layer menu lets the user switch layers on and off for each individual goblet to allow immediate comparisons. Selecting a place displays a popup with a textual transcription from the Goblets, and metadata drawn from the gazetteer. What is noteworthy about this mashup, however, is not so much the map itself – for which comparable projects already exist – but rather that the map can be automatically generated from a simple Pelagios data file, containing item metadata and annotations in Open Annotation RDF format. Thus the pathway from data production to visualization is both efficient and highly scalable across large numbers of documents.

Extracting places from maps: Ptolemy’s *Geographike Hyphegesis*

Previous work on toponym recognition in scanned maps focuses on contemporary documents for the simple reason that old maps remain extremely difficult for machines to parse. Our proposal is to automate the identification of potential toponyms in terms of their *location*, *extent* and *orientation* on the map image, so that researchers can then associate the results with items in pre-existing gazetteer lists and ultimately with URI-based gazetteers. The example given here is of Ptolemy’s regional map of Ireland and Great Britain (Fig. 1), digitized by the British Library.

Our first processing phase generates a black-and-white mask image, which isolates and separates “background” from “foreground”. The next phase locates and characterises features – in our case, connected objects – on the foreground image using an algorithm that detects contours. Since toponyms often consist of multiple features, the final phase aims to connect the detected features to groups that most likely represent a single toponym. Fig. 2 shows that for our test case the algorithm detected toponyms with a high success rate, correctly locating 38 of 41 places.

Our initial work with additional (including visually more complex) maps has raised several error scenarios and prompted some initial responses:

- **Ornament irritation.** Symbols and decorative elements that have structures in size and density (and colour) similar to toponyms frequently cause false positive detections. We expect that heuristics concerning the spatial density of matches and amount of overlap between them may be able to alleviate this problem, as these false detections exhibit distinctive clustering behaviour.
- **Line bleed.** Toponyms that intersect or are located near lines (often borders, graticules or rhumbs), can distort the recognition result. We expect that proper tuning of image processing parameters in the first separation step (such as colour thresholds, or thresholds determining the behaviour of line removal algorithms) may be able to lower the number of such errors, but it is unlikely that they can be avoided altogether. Increasing the efficiency of human verification and correction is essential for addressing this challenge.
- **Toponym crosstalk.** Especially in the presence of distracting elements such as lines, our algorithm can erroneously lead to toponym bounds that run across two neighbouring toponyms.. As in the case of errors caused by

line bleed, it is unlikely that these can be avoided, but metrics based on the morphology of the toponym may help to detect and flag them to a human operator for verification.

- **Split toponyms.** Our current processing approach does not specifically deal with toponyms that are split across multiple lines. An example can be found in Fig. 2, where "Alvion Insula Britannica" is split into two separate toponyms. Once again however, morphology and the spatial proximity of features will allow us to present human operators with potential candidates for merging into single features.
- **Large area & curvilinear toponyms.** Likewise, our heuristics are ill-suited to detect toponyms that cover large areas (e.g. regional toponyms), which are oriented significantly differently from other toponyms on the map, or which run along a curved baseline. Here, we may require a human to explicitly demarcate their bounds, although fortunately the size of such toponyms usually restricts their frequency in a given document.

While we expect that the amount of manual tuning and intervention will be further reduced by refining the processing workflow, toponym identification on old maps will never be a fully automated process. Therefore, we are also developing user interfaces and graphical tools that help both professional and public users carry out the manual work of aligning imagery to the gazetteers. On the one hand, we are experimenting with ways of re-presenting the user with re-oriented and visually enhanced image fragments so that they can be more easily interpreted. On the other we can use the spatial information in such images, and data from previous annotations, to propose likely candidates for 'one-click' annotations, and auto-completion of transcriptions.

Concluding Remarks

The data produced will provide us with opportunities to visualise both maps and texts in new ways. For instance, corpora of structurally similar documents, such as portolan charts, can be directly compared in terms of the places they refer to, the toponyms used, and their sequence along a coastline, in a similar manner to the itineraries described above. Alternatively, we can also blend out and replace toponyms with either modern or ancient alternatives where known, helping make these important documents easier to interpret for both scholars and public alike (Fig. 3). Most importantly we see this as the first steps in drawing new connections between the extraordinarily diverse range of early geospatial documents that have come down to us.



Fig. 1: Mashup showing route of the Vicarello Goblets (<http://pelagios.github.io/demos/vicarello-alpha/>)



Fig. 2: Part of Ptolemy's regional map of Ireland and Great Britain. (Ca. 1480 © The British Library Board. Harley MS 7182 ff. 60v-61.) Toponyms identified automatically and annotated with oriented bounding boxes.



Fig. 3: Original map (left) and map with original toponyms dynamically blurred out and replaced with corresponding modern place names.

References

- Rainer Simon, Elton Barker, Leif Isaksen.** (2012). *Exploring Pelagios: A Visual Browser for Geo-Tagged Datasets*. In International Workshop on Supporting Users' Exploration of Digital Libraries. Eneko Agirre, Kate Fernie, Arantxa Otegi, Mark Stevenson (Eds.) Cyprus, Paphos, September 27, 2012, pp. 29 - 34.
- Schmidt, M. G.** (2011), *A Gadibus Romam: Myth and Reality of an Ancient Route*. Bulletin of the Institute of Classical Studies, 54: 71–86.
- Elliott, T. & Gillies, S.** (2011). *Pleiades: An UnGIS for Ancient Geography*. Poster presented at Digital Humanities 2011, Stanford University. Available at: <http://dh2011abstracts.stanford.edu/xtf/view?docId=tei/ab-192.xml>
- Grover, C, Tobin, R, Byrne, K, Woollard, M, Reid, J, Dunn, S & Ball, J** (2010). 'Use of the Edinburgh geoparser for georeferencing digitized historical collections' In Philosophical Transactions of the Royal Society A: Mathematical, Physical & Engineering Sciences, vol 368, no. 1925, pp. 3875-3889. <http://nlp.stanford.edu/downloads/lex-parser.shtml>
See, for example, <http://vici.org/>
<http://www.openannotation.org/>
<http://www.bl.uk/onlinegallery/onlineex/unvbrit/p001hrl000007182u00060vrh.html>
- Rainer Simon, Peter Pilgerstorfer, Leif Isaksen, Elton Barker.** (2013). *Towards Semi-Automatic Annotation of Toponyms on Old Maps*. In 8th International Workshop on Digital Approaches in Cartographic Heritage. Rome, Italy, September 19-20, 2013.

Towards an Archaeology of Text Analysis Tools

Sinclair, Stéfan
McGill University, Canada

Rockwell, Geoffrey

University of Alberta, Canada

How have text analysis tools in the humanities been imagined in the past? What did humanities computing developers think they were addressing with now dated technologies like punch cards, printed concordances and verbose command languages? Whether the analytic functionality is at the surface, as with Voyant Tools, or embedded at deeper levels, as with the Lucene-powered searching and browsing capabilities of the Old Bailey, the web-based text analysis tools that we use today are very different from the first tentative technologies developed by computing humanists. Following Siegfried Zielinski's exploration of forgotten media technologies, this paper will look at three forgotten text analysis technologies and how they were introduced by their developers at the time. Specifically we will:

- Discuss why it is important to recover forgotten tools and the discourse around these instruments,
- Look at how punch cards were used in Roberto Busa's Index Thomisticus project as a way of understanding data entry,
- Look at Glickman's ideas about custom card output from PRORA, as a way of recovering the importance of output,
- Discuss the command language developed by John Smith for interacting with ARRAS, and
- Conclude with a more general call for digital humanities archaeology.

Zielinski and Media Archaeology

Siegfried Zielinski, in *Deep Time of the Media*, argues that technology does not evolve smoothly and that we therefore need to look at periods of intense development and then look at the dead ends that get overlooked to understand the history of media technology. In particular he shows how important it is to look at technologies that are not in canonical histories as precursors to "successful" technologies, because they provide insight into the thinking at the time. A study of forgotten technologies can help us understand opportunities and challenges as they were perceived at the time and on their own terms rather than imposing our prejudices. From the 1950s until the early 1990s there was just such a period of technology development around mainframe and personal computer text analysis tools. The tools developed, the challenges they addressed, and the debates around these technologies have largely been forgotten in an age of web-mediated digital humanities. For this reason we recover three important mainframe projects that can help us understand how differently data entry, output and interaction were thought through before born-digital content, output to wall-sized screens, and interaction on a touchscreen.

Busa and Tasman on Literary Data Processing

The first case study we will present is about the methods that Father Busa and his collaborator Paul Tasman developed for the *Index Thomisticus* (Busa could hardly be considered a forgotten figure, but he's often referred to metonymically as a founder of the field, with relatively little attention paid to the specifics of his work and his collaborations). Busa, when reflecting back on the project justified his technical approach as supporting a philological method of research aimed at recapturing the way a past author used words, much as we want to recapture past development. He argued in 1980 that, "The reader should not simply attach to the words he reads the significance they have in his mind, but should try to find out what significance they had in the writer's mind." (Busa 1980, p. 83) Concordances could help redirect readers towards the "verbal system of an author" or how the author used words in their time and away from the temptation to interpret the text at hand using contemporary conceptual categories. Concordancing creates a new text that shows the verbal system, not the doctrine.

Busa's collaborator Paul Tasman, however, presents a much more prosaic picture of their methodology that focuses on data entry using punch cards so you can actually get concordances

of words. He published a paper in 1957 on "Literary Data Processing" in the *IBM Journal of Research and Development* that focuses on how they prepared their texts accounting for human error and other problems. Tasman writes, "It is evident, of course, that the transcription of the documents in these other fields necessitates special sets of ground rules and codes in order to provide for information retrieval, and the results will depend entirely upon the degree and refinement of coding and the variety of cross referencing desired." (p. 256) This case study takes us back to a forgotten set of problems (representing text using punch cards) which led to more mature issues in text encoding. In the full presentation we will look closely at the data entry challenges faced by Busa's team and how they were resolved with the card technology of the time.

Glickman and Stallman on Printed Interfaces

The second case study we will look at is the development of the PRORA programs at the University of Toronto in the 1960s. PRORA was reviewed in the first issue of CHUM and with the publication of the *Manual for the Printing of Literary Texts and Concordances by Computer* by the University of Toronto Press in 1966 is one of the first academic analytical tools to be formally published in some fashion. What is particularly interesting, for our purposes, is the discussion in the Manual of how concordances might be printed. Glickman had idiosyncratic ideas about how concordances could be printed as cards for 2-ring binders so that they could be taken out and arranged on a table by users. He was combining binder technology with computing to reimagine the concordance text. Today we no longer think about output to paper as important to tools, and yet that is what the early tools were designed to do as they were not interactive. We will use this case study to recover what at the time was one of the most important features of a concording tool – how it could output something that could be published for others to use.

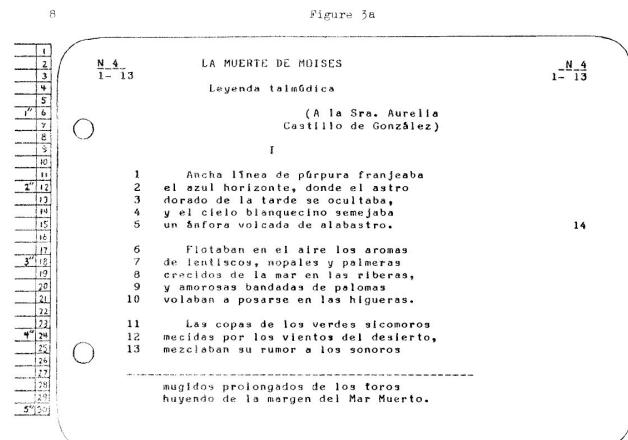


Fig. 1: Example of PRORA output from the Manual

Smith and Interaction

One of the first text analysis tools designed to support interactive research was John Smith's ARRAS. In ARRAS Smith developed a number of ideas about analysis that we now take for granted. ARRAS was interactive in the sense that it was not a batch program that you ran for output. It could generate visualizations and it was explicitly designed to be part of a multi-tasking research environment where you might be switching back and forth between analysis and word processing. Many of these ideas influenced the interactive PC concordancing tools that followed like TACT. In this paper, however, we are not going to focus on all the prescient features of ARRAS, but look at the now rather dated command language which Smith was so proud of. Almost no one uses a command language for text analysis any more; we expect our tools to have graphical user interfaces that provide affordances for direct manipulation. If you need to do something more than what Voyant, Tableau,

Lucene, Gephi or Weka let you do, then you learn to program in a language like R or Python. John Smith by contrast, spent a lot of time trying to design a natural command language for ARRAS that humanists would find easy to use and this comes through in his publications on the tool (1984 & 1985). Command languages were, for a while, the way you interacted with such systems and attention to their design could make a difference. Smith tried to develop a command language that was conversational so humanists could learn to use it to explore “vast continents of literature or history or other realms of information, much as our ancestors explored new lands.” (Smith 1984, p. 31) Close commanding for distant reading.

Conclusions

In the 2013 Busa Award lecture Willard McCarty called us to look to our history and specifically to look at the “incunabular” years before the web when humanists and artists were imagining what could be done. One challenge we face in reanimating this history is that so much of the story is in tools, standards and web sites – instruments difficult to interrogate the way we do texts. This paper looks back at one major thread of development - text analysis tools – not for the entertainment of outdated technology, but recover a way of thinking about technology. We will conclude by discussing other ways back including the need for better documentation about past tools, along the lines of what TAPoR 2.0 is supporting, and the need to preserve tools or at least a record of their usage.

References

- Busa, R.** (1980). "The Annals of Humanities Computing: The Index Thomisticus." *Computers and the Humanities*. 14(2): 83-90.
- Glickman, Robert Jay, and Gerrit Joseph Staalman.** *Manual for the Printing of Literary Texts and Concordances by Computer*. Toronto: University of Toronto Press, 1966.
- Liu, Alan.** (2012) "Where is Cultural Criticism in the Digital Humanities." In Debates in the Digital Humanities. Ed. Matthew K. Gold. University of Minnesota Press. Liu's essay is online at <<http://dhdebates.gc.cuny.edu/debates/part/11>>.
- Smith, J. B.** (1978). "Computer Criticism." *STYLE* XII(4): 326-356.
- Smith, J. B.** (1984). "A New Environment For Literary Analysis." *Perspectives in Computing* 4(2/3): 20-31.
- Smith, J. B.** (1985). *Arras User's Manual: TR85-036*. Chapel Hill, NC, The University of North Carolina at Chapel Hill.
- Tasman, P.** (1957). "Literary Data Processing." *IBM Journal of Research and Development* 1(3): 249-256.
- Zielinski, Siegfried.** (2008) *Deep Time of the Media: Toward an Archaeology of Hearing and Seeing by Technical Means*. Cambridge, Massachusetts: The MIT Press.

Advocating for a Digital Humanities Curriculum: Design and Implementation

Smith, David
david.chan.smith@gmail.com
 Wilfrid Laurier

Perceptions of falling enrollments and demands for closer alignment with the labour market have placed the humanities under pressure in North American higher education.[1] These issues have been particularly pressing at mid-sized institutions such as Wilfrid Laurier University in Ontario, Canada.[2] Faculty at Laurier responded to this challenge by developing a digital humanities program with its first course offerings expected in 2014-2015. This paper discusses the initial design of that program and its relationship to three major questions in digital humanities pedagogy. First, should digital humanities

programs be structured around a common core of learning objectives, or instead differentiate?[3] Second, what is the relationship between digital humanities curricula and demand in the workforce – should digital humanities programs be designed to pursue indifferent academic knowledge or attempt to engage more actively with vocational preparation? Finally, should the teaching of digital humanities focus on specific skill development, or instead cultivate “methodologies” or critical perspectives on technology and its application?[4]

In addressing these questions, the paper adds the experience of one institution to the ongoing conversation among emerging programs.[5] Although it has been asked whether only research intensive universities might have the expertise to field digital humanities programs, at Laurier we were mindful of the possibilities opened by a primarily pedagogical approach. [6] To begin with, while the secondary literature’s discussion of digital humanities research overshadows teaching, pedagogical arguments can be crucial to attract administrative support. [7] This is especially the case in faculties that are sensitive to undergraduate enrollments and their volatility. Digital humanities programs can present an attractive option for students while demonstrating the continuing relevance of the humanities to technological change.[8]

The Laurier curriculum is a program option that majors or minors can include in their course of study. Structured around a cluster of required courses, the curriculum allows departments throughout the Faculty of Arts to add courses as they are developed. Students use their elective courses to design a specific pathway, including learning to program.[9] The required courses expose students to historical analysis and computer science. The purpose of the curriculum, broadly conceived, is to improve students’ “digital literacy” or their ability to find and analyze digital information, and to use digital tools in active and creative ways.[10] Recent research has noted that, “There is considerable evidence to support the view that many students do not explore information in any deep or reflective manner.”[11] Other authors have preferred to emphasize student “multiliteracy,” arguing that literacy in the digital age is a broad concept, and reflects fluency with and access to a broad range of representative forms, such as visual or audio media.[12] A narrow concentration on traditional textual literacy, it is argued, misses the scope of literacy in a connected, technologically saturated world.[13] Though some commentators also worry that young people are being transformed into passive recipients of digital media, others argue that technology opens their creative potential in blogs and other formats.[14] All agree, however, with the necessity of transforming students from “consumers” of digital content into “creators.”[15] The literature has also argued that digital humanities can develop students’ critical skills as they engage with complex digital information on the web and elsewhere.[16]

Mindful of the growing significance of digital literacy, the Laurier program prompts students to realize the challenge of using and gathering “deep data,” rather than relying on data returned from basic Google searches. Throughout the curriculum students use methods from history to interpret textual information, and weigh and contextualize evidence. This approach connects a qualitative layer to the quantitative and analytic skills learned from computer science.[17] For example, the program’s foundation course introduces students to the possibilities of big data. Using existing queries and code they investigate familiar data sources, such as Twitter and Google. They then use knowledge from the humanities to contextualize and shape that data. Students are asked to consider the limitations of digital information and how to make data meaningful: what socially significant questions might they ask of it? During the final project students communicate their findings using digital media. The course attempts to demonstrate to students that the familiar digital universe they inhabit can reveal surprising discoveries with the right tools.

At Laurier three factors shaped the development of the program: concerns over costs, an increasing emphasis on differentiation within the Ontario university system, and the challenge of engaging faculty who had a pre-existing knowledge of the subject. These pressures demanded the program leverage existing institutional strengths.[18] For example, without funds to support new hires, the program

was by necessity an interdisciplinary effort among the faculty already working in the digital humanities. Consequently, their knowledge directly affected what was initially possible within the program.[19] As we developed the program we realized that it was possible for these perceived drawbacks, such as lack of faculty expertise concentrated in a research cluster, to become strengths. In response, our program became not only interdisciplinary, but a scaffold for faculty to build their expertise and advance their knowledge through teaching.

These factors shaped the curriculum so as to differentiate it within the Ontario system at a time when such diversity is becoming a compelling trend in higher education. As universities attempt to communicate their distinctiveness to applicants, digital humanities programs can benefit by their alignment with the institution's academic identities.[20] In the case of Laurier, this tilted the curriculum towards business, one of the university's strongest areas. Our experience suggests explicit differentiation is not only the preferable strategy, but also perhaps a necessity given the resource constraints and the dynamics of higher education in North America.

The development of the Laurier program was also related to specific data about the job-market. Whether humanities programs should explicitly adopt a vocational orientation has been a subject of pyretic debate.[21] A curriculum that trains students primarily to investigate academic problems in the humanities might be especially suitable for research universities. At Laurier, however, we shaped the curriculum in consultation with the Career Centre to advance our students in post-graduate employment more directly. Like it or not, many students and applicants are preoccupied with the job prospects associated with their major.[22] Among employers, we have learned, there is concern that graduates in arts will be intimidated or flummoxed by even basic tasks using digital tools. These misgivings might be unfounded, but the Laurier program explicitly cultivates digital literacy to equip students for knowledge employment in the future.

For example, the program builds on Laurier's strength in business administration to provide an entry point into the burgeoning field of analytics. This focus is especially important since the university is located in a region with large technology and insurance sectors.[23] The curriculum exposes students to big data problems beginning in the foundation course, while prompting them to think about the social meaning and application of this information by drawing on knowledge from the humanities. A stream may be added to educate undergraduates specifically in big data and analytics.

An emphasis on employable experience is also reflected in the experiential and co-op learning integrated into the curriculum.[24] Students, to give one example, can receive course credit for work at the Laurier Centre for Military Strategic and Disarmament Studies. They will undertake projects such as digitizing and making the Centre's archival holdings publicly searchable. These kinds of work opportunities, which combine training in the humanities with digital work, are not only pedagogically desirable, but meet our students' demand for co-op experience in the humanities.

Among the hardest questions to answer is whether digital humanities courses should teach a defined set of skills over more broadly conceived methodologies. In balance is the preference among employers that new hires should already have a minimal level of job-related training, yet programs focused on imparting specific skills risk narrowness.[25] Instead we have embraced the concept of digital literacy: students should have broad facility with digital work, and be confident and able to self-learn or advance their training in specific areas.[26] For example, the exposure during the foundation course to code does not teach them coding, but rather demonstrates how code works and its limitations. Students choosing to specialize can take advanced programming electives. However, all students should leave the program with at least enough understanding to customize off-the-shelf tools.

Can the digital humanities draw attention to the vitality of the humanities? Part of doing so may be to demonstrate that the humanities have much to offer the digital economy. Moving forward at Laurier, we intend to conduct a more thorough investigation into the relationship between the digital humanities and the contemporary workplace. By proceeding with this

research we accept that the humanities can be more explicitly oriented to post-graduate employment and the challenges of "knowledge work." Though we live in a time of doubt about the humanities, such strategies may instead reveal that this is a moment of renewed vigor.

References

- [1] Jennifer Levitz and Douglas Belkin, *Humanities Fall from Favor*, The Wall Street Journal (June 6, 2013), A3; Tamar Lewin, *As Interest Fades in the Humanities, Colleges Worry*, The New York Times (October 31, 2013), A1. For an articulate discussion of the falling numbers of humanities concentrators at a leading research university see, David Armitage, Homi Bhabha et al., *The Teaching of the Arts and Humanities at Harvard College: Mapping the Future*, (Cambridge, MA, 2013); and the analysis of Anthony Grafton and James Grossman, *The Humanities in Dubious Battle*, The Chronicle of Higher Education, July 1, 2013; Contrarian views are offered by Michael Bérubé, *The Humanities, Declining? Not According to the Numbers*, The Chronicle of Higher Education, July 1, 2013; and Robert Townshend, *Clio's Charm Holding Fast? Perspectives on History* (October, 2012).
- [2] The author's home institution; its full-time undergraduate population is approximately 15,400.
- [3] Lisa Spiro, *Opening up Digital Humanities Education*, in Digital Humanities Pedagogy, pp. 338-39.
- [4] Simon Mahony and Elena Pierazzo, *Teaching Skills or Teaching Methodology?* in Brett Hirsch (ed.), Digital Humanities Pedagogy: Practices, Principles and Politics (2012), pp. 215-225; at p. 215.
- [5] See, for example, the informal survey of Tanya Clement, *Multiliteracies in the Undergraduate Digital Humanities Curriculum: Skills, Principles, and Habits of Mind*, in Digital Humanities Pedagogy, pp. 365-88, at pp. 376-384.
- [6] Bryan Alexander and Rebecca Frost Davis discuss the challenges of digital humanities education in liberal arts schools and the resource limitations in *Should Liberal Arts Campuses Do Digital Humanities?* in Matthew K. Gold, Debates in the Digital Humanities (Minneapolis, 2012), pp. 368-89.
- [7] Tanya Clement, *Multiliteracies in the Undergraduate Digital Humanities Curriculum*, pp. 366, 370-71.
- [8] Luke Waltzer, *Digital Humanities and the 'Ugly Stepchildren,' of American Higher Education*, in Gold, Debates, pp. 335-349, 341-42; Stephen Brier, *Where's the Pedagogy? The Role of Teaching and Learning in the Digital Humanities*, in Gold, Debates, pp. 390-401, at 390-1; and Alan Liu, *Digital Humanities and Academic Change*, English Language Notes 47 (2009), pp. 17-35. Recent discussion of undergraduate pedagogy includes, T. Mills Kelly, *Teaching History in the Digital Age* (Ann Arbor, 2013) and throughout Brett Hirsch (ed.), Digital Humanities Pedagogy. Attempts have been made to identify the learning outcomes of digital humanities courses, see the report of Joanna Drucker and John Unsworth on the NEH Digital Humanities Curriculum Seminar at jefferson.village.virginia.edu/hcs/dhcs/intro_syllabus.html.
- [9] The role and necessity of coding skills in digital humanities program has been a subject of recent discussion, see *Interchange: The Promise of Digital History*, The Journal of American History 95:2 (Sept. 2008), pp. 452-487, at pp. 459-67; Tanya Clement, *Multiliteracies in the Undergraduate Digital Humanities Curriculum*, p. 369; Stephen Ramsay, *Programming with Humanists: Reflections on Raising an Army of Hacker-Scholars in the Digital Humanities*, in Hirsch (ed.), Digital Humanities Pedagogy, pp. 227-239.
- [10] Tanya Clement, *Multiliteracies in the Undergraduate Digital Humanities Curriculum*, p. 366. A recent report in the United Kingdom by a panel of higher education administrators and faculty defined "information literacy" as "activities such as search, retrieval and critical evaluation information from a range of sources, and also its responsible use form the point of view of attribution." David Melville et al., *Higher Education in a Web 2.0 World* (March, 2009), p. 34.
- [11] Peter Williams and Ian Rowlands, *The Literature on Young People and their Information Behaviour, work package II* (October, 2007), pp. 17-20, at 19. Other writers have gone

further, suggesting that the very process of web interaction and the structure of its information negatively affects users, **Nicholas Carr**, *Is Google Making us Stupid? What the Internet is Doing to Our Brains*, The Atlantic (July/August, 2008). **Daniel Cohen and Roy Rosenzweig** examine historical accuracy on the web in *Web of Lies? Historical Knowledge on the Internet*, First Monday (December, 2005). See also **Mark Bauerlein**, *The Dumbest Generation: How the Digital Age Stupifies Young Americans and Jeopardizes Our Future* (2008). See also the National Endowment for the Arts, *Reading at Risk: A Survey of Literary Reading in America*, Research Division Report #46, eds. **Tom Bradshaw and Bonnie Nichols** (Washington, 2004), eds.

[12] **New London Group**, *A Pedagogy of Multiliteracies: Designing Social Futures*, Harvard Educational Review 66:1 (1996), pp. 60-92.

[13] Ibid., p. 61.

[14] **Tanya Clement**, *Multiliteracies in the Undergraduate Digital Humanities Curriculum*, pp. 375-76.

[15] **Melville** et al., *Higher Education in a Web 2.0 World*, pp. 22-3.

[16] **Tanya Clement**, *Multiliteracies in the Undergraduate Digital Humanities Curriculum*, p. 366.

[17] The program also prompts students to reflect on the use of large-scale historical data, issues that have been explored by **William Thomas III**, *Computing and the Historical Imagination*, in Susan Schreibman, Ray Siemens et al. (eds.), *A Companion to Digital Humanities*, (Oxford, 2004), pp. 56-68, esp. p. 65 for a discussion of the future digital needs of historians. **Frederick Gibbs and Trevor Owens** also suggest that the availability of "big data" requires the historical profession to rethink its methods and hermeneutics, see *The Hermeneutics of Data and Historical Writing*.

[18] **Melissa Terras**, **Julianne Nyhan** and others have recently compiled differing perspectives on the digital humanities in *Defining Digital Humanities: a Reader* (Surrey, UK, 2013).

[19] Undergraduate programs can be capacity building by attracting resources to digital humanities. **Spiro**, *Opening up Digital Humanities Education*, p. 332.

[20] **James Bradshaw**, *Specialize or Risk Losing Funding Ontario tells Universities and Colleges*, The Globe and Mail, September 18, 2013.

[21] In the United Kingdom, Jisc has sponsored projects considering the linkages between digital literacy and employment through the Developing Digital Literacies Programme. Examples include the University of Greenwich Digital Literacy in Higher Education Project, and the Digitally Ready project at the University of Reading.

[22] The recent Gallup-Lumina survey reported that 90% of respondents among the general public believed that the "candidates college or university major" was either "somewhat" or "very important" to hiring decisions. Only 70% of "business leaders" who were asked responded similarly. The 2013 Lumina Study of the American Public's Opinion on Higher Education and U.S. Business Leaders Poll on Higher Education (February 25, 2014), pp. 18, 29

[23] **Andrew McAfee and Erik Brynjolfsson**, *Big Data: The Management Revolution*, Harvard Business Review (October, 2012), p. 62; **Jonathan Shaw**, *Why 'Big Data' is a Big Deal: Information Science Promises to Change the World*, Harvard Magazine (March-April, 2014), pp. 30-35; 74-75; **Claire Miller**, *Data Science: The Numbers of Our Lives*, The New York Times, April 11, 2013; **Thomas Davenport and D.J. Patil**, *Data Scientist: the Sexiest Job of the 21st Century*, Harvard Business Review (October, 2012), pp. 70-1; **Alexandra Stevenson**, *New Silicon Valley Fund to Back Big Data Start-Ups*, New York Times Dealbook, October 17, 2013. **Claire Miller**, *Data Science: The Numbers of Our Lives*, The New York Times , April 11, 2013.

[24] 14% of business leaders surveyed in 2014 reported that internships or practical experience would "best prepare graduates for success in the workforce." This was the most popular response (excluding "Don't know/refused"). The 2013 Lumina Study, p. 30. For discussion of such "participatory" educational opportunities (or their lack in the humanities) see **Alexander and Frost Davis**, *Liberal Arts Campuses*, p. 380; **Spiro**, *Opening Up Digital Humanities Pedagogy*, p. 352; **Cathy**

Davidson has called for a "core curriculum to create engaged entrepreneurs," here.

[25] Skills may also, without use, atrophy unlike patterns of critical thinking, see **Mahony and Pierazzo**, *Teaching Skills or Teaching Methodology?* p. 224.

[26] **Jisc**, *Learning Literacies in a Digital Age*, September 7, 2009.

Transcriptional implicature: a contribution to markup semantics

Sperberg-McQueen, Michael

cmsgmcq@blackmesatech.com
Black Mesa Technologies LLC, United States of America

Marcoux, Yves

ymarcoux@gmail.com
Université de Montréal

Huitfeldt, Claus

Claus.Huitfeldt@uib.no
Universitetet i Bergen

We may see, in a TEI transcription of an old book, the lines:

<pb n="["iii"]"/> <p>Quaestiones, quae ad mathematicae fundamenta pertinent, etsi hisce temporibus a multis tractatae, satisfaciens solutione et adhuc carent.

What do they mean? How do we know what they mean? Can we model their meaning formally?

Formalizing the meaning of arbitrary natural-language utterances remains intractable today. But markup languages, being formally defined artificial languages, appear more approachable. So we may be able to explain what the <pb> and <p> elements mean, even if the sense of the Latin eludes formalization. Some propose to explicate the meaning of markup by specifying, for each construct in a markup vocabulary, a sentence schema in a natural language, with blanks to be filled in with data from the document¹; others make a similar proposal but allow sentence schemata in formal languages like first-order predicate logic as well². This appears straightforward, although far from trivial, for metadata³ and perhaps even for born-digital texts, but how shall the meaning of <p> be formalized in a markup language which defines it as containing a *transcription of a text block in a manuscript*? What does it mean for a document to be a transcription of another document? Can we formalize that?

Earlier work has explored the nature of the similarity between transcripts and their exemplars. Perhaps it consists simply in their containing the same sequence of characters? This can be formalized but proves disappointing, partly because it omits text structures like division into paragraphs and partly because it offers no way of describing disagreements among transcribers about how to read the exemplar, or which character distinctions (e.g. i/j, u/v, s/l) to retain and which to level. It is also wrong: few transcripts have exactly the same character sequence as their exemplar⁴. Later work extends the analysis from characters to higher-level textual structures and models transcriber agreement and disagreement explicitly^{5 6}. (Paul Caton has built on this idea to propose *pure transcriptional markup* as an approach to the problems of transcription semantics⁷.) By treating higher-level constructs as tokens of higher-level types and using the sentence schemata mentioned earlier, we believe one can formalize the meaning of tokens (characters, words, and XML elements like <p> and <pb> alike) in transcriptions.

Clearly, however, the meaning of tokens in a transcription depends on the transcription conventions adopted, which vary. There is hardly any universal transcription practice: for every generalization we find exceptions^{8 9 10 11}. Is everything in the exemplar transcribed? Not when deletions and irrelevant material are excluded. Does everything in the transcription reproduce some word or character in the exemplar? Not when line breaks are marked explicitly with vertical bars, or notes

are added. Many scholarly editions account for variations like these in an explicit statement of transcription practice. Such statements typically describe ways in which practice varies from the usual practice, but rarely the ways in which it exemplifies normal practice. In any community of scholarship, some common practice is typically felt to be so obvious that it needs no mention or explanation.

One job of formalization is to make explicit practices and assumptions otherwise passed over in silence.

We propose a notion we shall call *transcriptional implicature*, denoting a set of rules which apply by default but which may be overridden in particular cases, analogous to the rules of conversational implicature proposed by H. P. Grice as a way of explicating the logic of everyday conversation¹². The operational definition of transcriptional implicature for a given community is “the set of rules no one in the community bothers to mention explicitly”.

Different communities of transcription practice have different sets of tacit assumptions and thus different rules of transcriptional implicature. Is there a common core of transcriptional practice shared by all communities? Maybe; it's an empirical question. A serious answer would require detailed study of a wide variety of communities of practice. We postulate, however, that the transcriptional implicature of any community of practice can be described with reference to some default set of rules for transcriptional implicature.

The transcriptional practice of any given project is commonly documented by listing its deviations from the transcriptional implicature of the relevant community. If that transcriptional implicature can (as postulated) be described as a set of deviations from the default transcriptional implicature, then it follows that any project's transcription practice can be described with reference to the default transcriptional implicature, by merging the two lists of differences.

We propose to identify this hypothetical *default transcriptional implicature* with the rules outlined below. In the formalizations, “*T*” denotes a transcript, “*E*” its exemplar.

Adopting the extended use of the type/token distinction mentioned above, the default transcriptional implicature can be summed up in a single rule:

1. 1. A transcript and its exemplar have the same type.

Formally:

$$\text{type}(T) = \text{type}(E)$$

In interesting cases, *E* will have a complex type consisting of some structure of smaller types (which in turn consist of smaller ones still), instantiated by a complex token which similarly consists of smaller tokens. It is a consequence of (1) that:

1. 2. There is a one-to-one correspondence between the tokens of a transcript and the tokens of its exemplar, such that every pair of corresponding tokens have the same type.

Formally, this is a second-order statement, but we can approximate it using the following first-order sentence, which assumes a function *tokens* which maps from a document to the set of tokens contained in that document, and the relations *RET* (mapping from *tokens(E)* to *tokens(T)*) and *RTE* (the other way round).

$$(\forall t1 : \text{tokens}(E)) (\exists t2 : \text{tokens}(T)) (t2 = \text{RET}(t1)) \wedge \\ (\forall t1 : \text{tokens}(T)) (\exists t2 : \text{tokens}(E)) (t2 = \text{RTE}(t1)) \wedge \\ (\forall t1 : \text{tokens}(E), t2 : \text{tokens}(T)) (t2 = \text{RET}(t1) \Leftrightarrow t1 = \text{RTE}(t2) \wedge \text{type}(t1) = \text{type}(\text{RET}(t1)))$$

It is easier to relate variations in transcription practices to the default transcriptional implicature if we paraphrase (2) as a conjunction of simpler rules (3) - (6):

1. 3. For every token in the exemplar there is exactly one corresponding token in the transcript.

$$(\forall t1 : \text{tokens}(E)) (\exists t2 : \text{tokens}(T)) (t2 = \text{RTE}(t1))$$

Applied to the example with which this document begins, this means: each token in the exemplar maps to a token in the transcript. We can infer that the sentence quoted does not contain the word *non*, because otherwise *non* would appear in the transcript.

2.

4. For every token in the transcript there is exactly one corresponding token in the exemplar.

$$(\forall t1 : \text{tokens}(T)) (\exists t2 : \text{tokens}(E)) (t2 = \text{RET}(t1))$$

Applied to the example: each token in the transcript maps to some token in the exemplar. We can infer that the exemplar contains some token corresponding to the word *Quaestiones* in the transcript.

3. 5. The relations identified in rules (3) and (4) are inverses: that is, for every pair of tokens *t1* in the exemplar and *t2* in the transcript, if *t2* corresponds to *t1* as described in rule (3), then *t1* corresponds to *t2* as described in rule (4).

$$(\forall t1 : \text{tokens}(E), t2 : \text{tokens}(T)) (t2 = \text{RET}(t1) \Leftrightarrow t1 = \text{RTE}(t2))$$

4. 6. In every pair of corresponding tokens, the two tokens are tokens of the same type.

$$(\forall t1 : \text{tokens}(E)) (\text{type}(t1) = \text{type}(\text{RET}(t1)))$$

Applied to the example: we can infer that the token in the exemplar which corresponds to the word *Quaestiones* in the transcript is itself a token of the same word type.

We observe that many, perhaps all, variations in transcription practice can be classified by which rule they override.

- Silent expansion of abbreviations and normalization of spelling exclude some individual characters in *E* and *T* from the scope of rules (3) and (4); those rules typically still apply to words and higher-level tokens.
- Expansion of abbreviations in brackets can preserve the character-level mapping, but introduces characters in *T* which are exceptions to rule (4), since they lack corresponding characters in *E*. The use of vertical bars (|) in *T* to record line breaks in *E* is also an exception to rule (4).
- Omission of selected material (deleted words, additions from a second hand, ...) modifies rule (3) by identifying tokens in *E* which are not represented in *T*.
- Both transcriptions which distinguish archaic allographs (i/j, u/v, s/l) and those which level non-graphemic distinctions obey rule (6), but they read the document with different type systems.
- The principle of charity (“in cases of doubt assume *E* is correct”) can also be interpreted as a further elaboration of rule (6).

The full paper will explore these and further ways in which the practice of transcription can deviate from the default rules of transcriptional implicature as we proposed to define them, and show how the variations in practice can be described formally.

References

1. Marcoux, Yves (2006). *A natural-language approach to modeling: Why is some XML so difficult to write?* Paper given at Extreme Markup Languages, Montréal. Proceedings of Extreme Markup Languages 2006. On the Web at conferences.idealliance.org/extreme/html/2006/Marcoux01/EML2006Marcoux01.html .
2. Sperberg-McQueen, C. M., Claus Huitfeldt, and Allen Renear (2001). *Meaning and interpretation of markup*. Markup Languages: Theory & Practice 2.3: 215–234. On the Web at cmsmcq.com/2000/mim.html
3. Wickett, Karen M., and Allen Renear (2009). *A first order theory of bibliographic objects*. Proceedings of the American Society for Information Science and Technology 46.1: 1–8. On the Web at onlinelibrary.wiley.com/doi/10.1002/meet.2009.1450460378/full (subscription required).
4. Huitfeldt, Claus, and C. M. Sperberg-McQueen (2008). *What is transcription?* Literary & Linguistic Computing 23.3: 295–310.
5. Sperberg-McQueen, C. M.. Claus Huitfeldt, and Yves Marcoux (2009). *What is transcription? Part 2*. Talk given at Digital Humanities, College Park, Maryland. Slides on the Web at blackmesatech.com/2009/06/dh2009 .
6. Huitfeldt, Claus, Yves Marcoux, and C. M. Sperberg-McQueen (2010). *Extension of the type/token distinction to document structure*. Paper presented at Balisage: The Markup Conference 2010, Montréal, Canada, August 3 - 6,

2010. In Proceedings of Balisage: The Markup Conference 2010. Balisage Series on Markup Technologies, vol. 5. doi:10.4242/BalisageVol5.Huitfeldt01. On the Web at www.balisage.net/Proceedings/vol5/html/Huitfeldt01/BalisageVol5-Huitfeldt01.html.
7. **Caton, Paul** (2013). *Pure transcriptional encoding*. Paper given at Digital Humanities 2013, Lincoln, Nebraska.
 8. **Carter, Clarence E.** *Historical editing. Bulletins of the national archives, Number 7*. [Washington, DC]: National Archives and Records Service, August 1952. National Archives publication number 53-4.
 9. **Tanselle, G. Thomas** (1989). *A Rationale of Textual Criticism* Philadelphia: University of Pennsylvania Press. 104 pp.
 10. **Vander Meulen, David, and G. Thomas Tanselle** (1999), *A system of manuscript transcription Studies in Bibliography* 52: 201-212.
 11. **Robinson, Peter, and Elizabeth Solopova** (2006), *Guidelines for Transcription of the Manuscripts of The Wife of Bath's Prologue*. 18 March 2006. On the Web at www.canterburytalesproject.org/pubs/transguide-M1.pdf.
 12. **Grice, H.P.** (1975). *Logic and Conversation*, Syntax and Semantics, vol.3 edited by P. Cole and J. Morgan, Academic Press. Reprinted as ch.2 of his *Studies in the Way of Words* (Cambridge, Mass.: Harvard University Press, 1989), pp. 22–40.

Arch-V: A Platform for Image-Based Search and Retrieval of Digital Archives

Stahmer, Carl
cstahmer@gmail.com
 UC Santa Barbara

1. Introduction

Arch-V (short for Archive Vision) is a newly developed C++ application that provides image based search capabilities for digital archives of print materials. In 2013 the English Broadside Ballad Archive (EBBA) at the Early Modern Center, University of California, Santa Barbara was awarded an NEH Start-Up Grant to begin work on the Ballad Impression Archive (BIA), a component of EBBA specifically devoted to cataloguing the over 9,000 (and growing) individual woodcut impressions in EBBA and making them fully searchable as an image collection. A key component of this award was the creation of software to provide automated, image based searching of the collection. The proposed short paper will introduce and provide an overview of the implementation procedures for Arch-V.

1.1. Overview

Arch-V is a platform for delivering automated, image-based indexing and searching of digital archives. While the state of the art in computerized image classification and recognition is quite advanced, the application of these technologies to the specific area of digital archives of printed material presents a unique set of challenges. As noted by Relja Arandjelovic and Andrew Zisserman, the problem of automated recognition of objects has been largely solved, “provided they have a light coating of texture”¹. This is because the state of the art in computer vision relies upon the refraction of light across the surface texture of an object as it is captured in a digital image (or frame of video) in order to extract recognizable feature points as indexable markers of the object in the image. But in digital images of print artifacts, surface texture serves as a distraction from and not an indicator of the objects depicted in the print. This is because the texture belongs to the delivery medium, the carrier, and not to the objects being represented.

As a result, the efficacy of current technologies is less than satisfying when applied to the area digital archives of printed materials. More importantly, this is not a problem that computer science researches are likely to solve for the humanities, as the primary interest, funding, and work effort in computer science is in the area of processing networked picture and video feeds such as surveillance footage, YouTube videos, and Facebook photos.

We were able to design and test a solution to this problem as part of the funding provided by the Start-Up award. This solution involves a process of normalizing color and black and white archival images to a common format prior to feature point extraction, utilizing a modified feature point extraction methodology, and combining the feature point extraction with a process of border contour extraction and comparison. This combination of practices allows us to produce a collection of feature points for each image that define the boundaries of the objects represented in them rather than variations in surface texture. Our solution has already been implemented in the EBBA cataloguing interface, and it will be implemented on the EBBA website in early 2014.

We continue to investigate and implement improvements (along lines identified during the start-up phase) to the image-based searching technology that specifically address its application for digital archives of print materials, to refactor the codebase as a distributable software package that can be easily implemented by other digital archives without advanced technical knowledge or experience, and to produce companion documentation to insure both success and ease of implementation by other archives. In its current form, the complete c++ codebase is publicly available via Git at <https://bitbucket.org/cstahmer/archv/>.

1.2. Methodology

Arch-V utilizes of novel combination of SURF feature point extraction of raw images, and feature point extraction of extracted contours from the base image set into order to create a Visual Dictionary of extant features. Each image in the archive is then processed using the same extraction algorithms and a Visual Word File representing a normalized histogram of the features found in each image is then created for each image. The Visual Word Files are then indexed using Lucene, which serves as the query engine for image comparison.

1.3 Scope of the Presentation

The proposed Short Paper will introduce the theoretical problems associated with performing visual searches of archives of print materials, give a short demonstration of the Arch-V software in action searching the over 9,000 images in the EBBA archive, and provide information on how users can implement Arch-V in their own archives.

References

1. **Reja Arandjelovic and Andrew Zisserman** (2011). *Smooth Object Retrieval using a Bag of Boundaries*. Proceedings of the IEEE International Conference on Computer Vision. www.robots.ox.ac.uk/~vgg/publications/2011/Arandjelovic11/arandjelovic11.pdf.

Digital Pedagogy is about Breaking Stuff

Stommel, Jesse
Jesse.Stommel@me.com
 University of Wisconsin-Madison, United States of America

John Dewey writes in *Schools of To-Morrow*: “Unless the mass of workers are to be blind cogs and pinions in the

apparatus they employ, they must have some understanding of the physical and social facts behind and ahead of the material and appliances with which they are dealing." This remark is not unlike the image Fritz Lang depicts at the outset of the 1927 film Metropolis: slaves to a machine becoming food for the machine. The danger in fetishizing machines is that we become subject to them. But turning away in the face of the digital will lead to much the same fate. Rather, we need to handle our technologies roughly -- to think critically about our tools, how we use them, and who has access to them.

Like Digital Humanities, Digital pedagogy has been variously defined. Brian Croxall and Adeline Koh offered a very inclusive, broad-stroke definition at their MLA Digital Pedagogy Unconference, saying that "digital pedagogy is the use of electronic elements to enhance or to change the experience of education." And Katherine D. Harris offered up the components of her digital pedagogy -- which she borrows in part from the "mainstays of Digital Humanities" -- during a NITLE seminar on the subject: "collaboration, playfulness/tinkering, focus on process, building (very broadly defined)."

Digital pedagogy is an orientation toward pedagogy that is not necessarily predicated on the use of digital tools. This is why I like Harris's focus on process and Croxall and Koh's use of the seemingly vague, but in fact quite lovely, phrase "electronic elements." The phrase dissects the notion of an educational technology, turning the discussion to a consideration of the smallest possible element that might influence teaching and learning: the electrical impulse. At this level, we're not talking about how we might use Wordpress in a composition class, or how Smart Boards failed to revolutionize K-12 education, but about how the most basic architecture of our interactions with and through machines can inspire new (digital or analog) pedagogies. Thus, Kathi Inman Berens says paradoxically that "the new learning is ancient."

Many have argued that the digital humanities is about building stuff and sharing stuff -- that the digital humanities reframes the work we do in the humanities as less consumptive and more curatorial, less solitary and more collaborative. I would argue, though, that the humanities have always been intensely interactive, an engaged dance between the text on a page and the ideas in our brains. The humanities have also always been intensely social, a vibrant ecosystem of shared, reworked, and retold stories. The margins of books as a vast network of playgrounds.

The digital brings different playgrounds and new kinds of interaction, and we must incessantly ask questions of it, disturbing the edge upon which we find ourselves so precariously perched. And what the digital asks of us is that every assumption we have be turned on its head. The digital humanities asks us to pervert our reading practices -- to read backwards, as well as forwards, to stubbornly not read, and to rethink how we approach learning in the digital age.

In fact, the course itself is one of our central texts, a collection of stories about reading and writing, that can be actively hacked and remixed. Sean Michael Morris writes, "A course today is an act of composition," an active present participle and not a static container. This is more and more true of courses that live online, which demand that we carefully examine the digital as a frame, while recognizing that the digital does not supersede and can never unseat the work we do in the world. Kathi Inman Berens writes, "It doesn't matter to me if my classroom is a little rectangle in a building or a little rectangle above my keyboard. Doors are rectangles; rectangles are portals. We walk through." This is where learning happens, at the breaking point of its various containers.

This is true just as well of the literary texts we analyze (and ask students to analyze) with digital tools. In the syllabus for a recent undergraduate seminar in the digital humanities, I pose the following questions:

- How is literature and our reading of it being changed by computers? What influence does the container for a text have on its content? To what degree does immersion in a text depend upon the physicality of its interface? How are evolving technologies (like the iPad) helping to enliven (or disengage us from) the materiality of literary texts?

Literature, film, and other media are changing, and the way we interact with them is also changing. As we imagine a digital approach to the humanities, we must look back even as we look forward, considering what media has become while we simultaneously examine the hows and whys of its becoming. We used to watch films only in a darkened theater without the distraction of other external physical stimuli. Increasingly, though, we watch film on hand-held digital devices, many with touch screens that allow more and more interaction with the content. Our apparatuses for media-consumption juxtapose digital media, literature, and film: Now, we watch Ridley Scott's Alien in a window alongside Twitter and Facebook. Film no longer exists as a medium distinct from these other media.

The same is true of new modes of reading. Digital texts invite (or allow) us to do other things with our eyes, brains, and bodies as we experience them. As I write this, I have 9 windows open on my computer, each vying for my attention. Some of these windows have several frames in further competition. Advertisements. E-mail. Documents. Widgets. Social-networking tools. Chat interfaces. Each of these layers has an effect on how I engage the digital text. In spite of all these layers, I don't think we experience a decreased attention; rather, the digital text demands a different sort of attention. Even as my direct engagement is challenged, my brain is offered more fuel for making connections and associative leaps. A proactive approach to online and digital pedagogy asks us to put these associative leaps to work. So, Twitter and FaceBook may be a distraction, but that distraction can be harnessed for good pedagogy.

Social media can function as a site for democratic participation, a leveled playing field, a harbinger for another sort of attention. The keenest analysis in the digital humanities is born of distraction and revels in tangents. The holy grail of this work is not the thesis but the fissure.

Breaking Stuff as an Act of Literary Criticism

The digital humanities is about breaking stuff. Especially at the undergraduate level, this is the work of the digital humanities that most needs doing. Mark Sample proposes "what is broken and twisted is also beautiful, and a bearer of knowledge. The Deformed Humanities is an origami crane -- a piece of paper contorted into an object of startling insight and beauty." And, by the end of a class, if it's successful, this is what becomes of the syllabus, the texts, the assignments, and us. Sample continues, "every fact is a fad and print is a prison. Instructors are insurgents and introductions are invasions." In this way, my digital humanities courses work to violently dismantle fact and print, instructors and introductions, and I revel together (and part and parcel) with students in both discovery and uncertainty.

The digital humanities course I teach for undergraduates has as its first assignment the breaking of something as an act of literary criticism. Specifically, I ask students to take the words of a poem by Emily Dickinson, "There's a certain slant of light," and rearrange them into something else. They use any or all of the words that appear in the poem as many or as few times as they want. What they build takes any shape: text, image, video, a poem, a pile, sense-making or otherwise.

This paper expands upon a brief article I wrote about this assignment, analyzing several of the resulting student works and exploring the new pedagogies that the digital humanities demand and give rise to.

References

- Dewey, John** (1915). *Schools of Tomorrow*. Montana: Kessinger Publishing.
- Dickinson, Emily** (1960). "There's a certain slant of light." *The Complete Poems*. Ed. Thomas H. Johnson. Boston: Little, Brown and Company.
- Harris, Katherine D.** (2012). "NITLE Digital Pedagogy Seminar." <http://triproftri.wordpress.com/2012/03/27/nitle-digital-pedagogy/> (accessed March 7, 2014).

- Inman Berens, Kathi** (2012). "The New Learning is Ancient." *New Media Curious*. <http://kathiiberens.com/2012/12/03/ancient/> (accessed March 7, 2014).
- Koh, Adeline and Brian Croxall** (2013). "What is Digital Pedagogy?" <http://www.briancroxall.net/digitalpedagogy/what-is-digital-pedagogy/> (accessed March 7, 2014).
- Morris, Sean Michael** (2012). "Courses, Composition, Hybridity." <http://www.seanmichaelmorris.com/courses-composition-hybridity/> (accessed March 7, 2014).
- Sample, Mark** (2012). "Notes Towards a Deformed Humanities." *Sample Reality*. <http://www.samplereality.com/2012/05/02/notes-towards-a-deformed-humanities/> (accessed March 7, 2014).

Creating a Digital Tombstone Archive: From Fieldwork to Theory Formation

Streiter, Oliver

ostreiter@nuk.edu.tw

National University of Kaohsiung, Taiwan

Goudin, Yoann

yoanngoudin@yahoo.fr

Institut National des Langues et Civilisations Orientales (INALCO), Paris, France

1. Introduction: A digital tombstone archive

1.1. Overview: Scope and motivation

Since 2007 we work on the construction and exploitation of a digital tombstone archive, "ThakBong", an archive which mainly contains tombstones of Taiwan, but includes also tombstones of China and tombstones of Chinese migrants in Asia, Europe and the USA. So far, on 850 visits to 500 graveyards, about 170.000 photos of 42.000 tombs have been taken. Repeated visits to graveyards allows us to document regionally different temporal patterns of ancestor veneration and the life-cycle of a tombs, including the burial, a second burial, a tomb renovation and the removal of the tomb. In fact, graveyards in Taiwan continue to disappear from the geographic and cultural landscape through development projects, natural catastrophes and the transformation of graveyards to bone-ash-towers. About half of all graveyards have already been lost or will be removed according to governmental projects in the years to come. As graveyards continue to be taboo for Taiwanese researchers, the records of graveyards we produce are among the only available.



Fig. 1: A graveyard in Southern Taiwan

Using geo-referenced photos as primary data, tombstones, tombs and graveyards are digitally reconstructed, linking back and forth between images and data. Images and data are constantly updated and made available to the scientific community via the research data archive DANS . They can be used in a wide range of research approaches, including corpus linguistics, social history, anthropology and human geography. The wealth of the data, and the interleaving levels of language, culture, geography and history however require the elaboration

of more integrative and cross-disciplinary approaches. Beyond the topics in relation to Taiwan, the data permit empirical tests for theories of colonization, globalization, cultural contagion , and social struggle through cultural practices . The approach that we take to our data is that of a Digital Anthropology that combines sociological theories with the transformation of cultural practices .

1.2. Scope of this Presentation

In this presentation we will summarize the main aspects of our digitization work, starting with the sampling and finishing with some theoretical insights we gained through our work. The purpose of this presentation is two-fold. First, we want to provide those who intend to launch a similar documentation and research project of their local graveyards with a short project description, which they might follow and, if needed, modify. Second, we want researcher who consider to use the "ThakBong" archive for their future studies to be able to evaluate the scope, coarseness and reliability of the data.

2. Methodology

2.1. Sampling

The purpose of the sampling is to produce an adequate representation of the reality in all dimensions we can think of, to avoid many of the unmotivated generalization we find in more traditional research. We are especially eager to capture the voices of poor and uneducated people, as they are usually overheard in linguistic and historic accounts. We therefore sample, without distinction, old and new tombs, king-size national monuments and most elementary tombs. In addition, we try to cover all ethnic groups, all religious orientations and all administrative divisions. For practical reasons, however, we cannot achieve a truly balanced sampling, because urban regions and catastrophe-prone areas have already lost most of their historic graveyards.

2.2. Fieldwork

The central tools for the fieldwork are digital GPS-cameras, costing about 250€, which store in the EXIF-header the position, altitude and orientation, along with more common metadata. For geographic analyses and the mapping of tombs, graveyards and cultural practices these data are indispensable.

To optimize the automatic use of these data, photos are taken in a regulated way, using two circles with two defined centers as a reference for the photos: All photos from outside are all taken into the direction of the tombstone, allowing the orientation of the tomb and the orientation of the shots to be calculated from the direction of one central camera shot. Photos in the area of the mourners, in front of the tombstone, are taken from the center of this space, allowing to calculate the location of the components of the tomb around the mourner from the orientation of the tomb. Also, through this model, photos are linked in a systematic way, making it possible to browse and virtually explore the tombs.

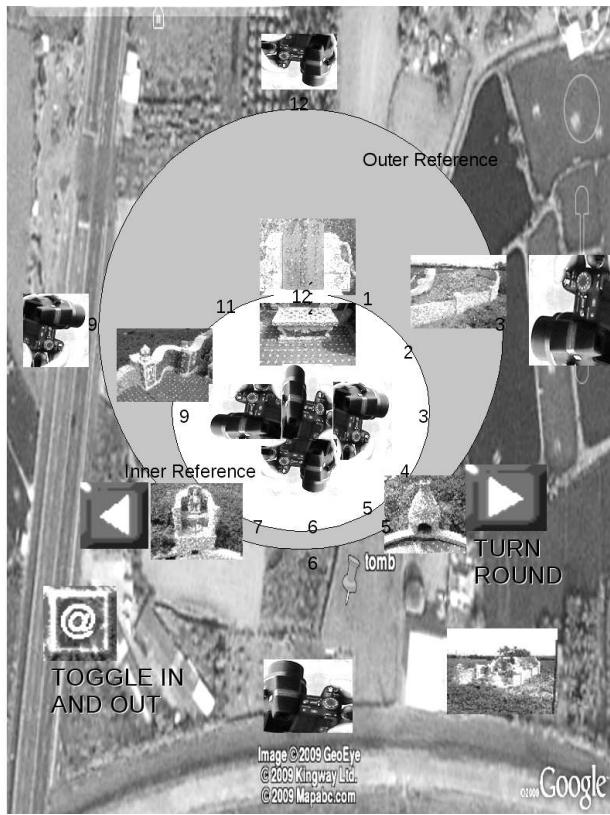


Fig. 2: The model regulating how photos are taken, so that automatic processing and browsing of the tomb become possible.

2.3. Processing of Primary Data

For the processing and multi-user annotation we use the PostgreSQL database the postGIS GIS extension, storing transcriptions in XML. The database tables represent digital objects as bundles of features. A *feature* is a unique combination of an attribute, e.g. 'direction', a *value*, e.g. '180' and if necessary a *unit*, e.g. 'degree'. These objects are called "graveyard", "tomb", "tombstone", "transcription" and "person", representing the corresponding real-world objects. The object "person", for example, contains the features 'surname', 'given name', 'religious name', 'date of birth', 'date of death', 'date of burial', 'ethnicity', 'gender' and 'role', e.g. 'mourner' or 'deceased'. An average tomb has about 50 defined features, but numbers might be much higher for family tombs.

In a first processing step, these objects are created manually through a web-interface that segments the stream of photos. Images that show an object are linked to that object. Images showing inscriptions, offerings, symbols and figurative representations are specifically tagged, so that images can be further filtered, to facilitate the annotation process.

More than 10.000 lines of program code in plpgsql implement PostgreSQL triggers, which help to reduce the manual annotation labor. They can be divided into five groups:

1. Rules that extract and processes data from the image. From these the position, altitude and slope are calculated.
2. Rules that fill in logically implied values. E.g. if no image of a tombs is tagged for 'offering', the tomb is marked as *offerings='no'*.
3. Rules for a model-based annotation. Models are used for non-visible features, such as the 'ethnicity' of the deceased, if not known otherwise. These statistical models maximize the number of correct annotations over the whole data set, allowing at the same time manual corrections of individual tombs to be respected and calculated into the model. Using external statistical data, for example, on the relation of surnames, administrative regions and ethnicity , the most likely ethnicity for the deceased can be calculated, given the region and the surname on the tombstone.
4. Where no external data resources are available, e.g. for the prediction of the 'gender' from the 'given name', we use

bootstrapping: Using statistical data that have been produced in the manual annotation of unambiguous cases, ambiguous cases are automatically annotated when the memory-based model makes clear predictions for a given name. For all features, their epistemic status are retained: Model-derived data are updated when manually set data change.

2.4. Annotation and Transcription

After experiments with OCR on tombstone images brought no results, tombstones are transcribed manually, phrase by phrase. Until now 24.000 tombstones have been completely, about 10.000 partially transcribed. Transcribed phrases are classified semi-automatically into 'semantic roles', such as 'place', 'person', 'date of birth' etc. Then, example-based taggers extract relevant data, such as dates, names and family relations and fill in the relevant features. Where automatic processes do not yield a clear classification, the system shifts to an interactive mode. All other features, e.g. the color and form of the tombstone are annotated manually until we will have completed the extraction of these features from other photos, using an example-based approach and a similarity metrics of photos. For the entire project, 6 man-years have been invested in 7 years, showing that with the right balance of automatic processing and manual annotation huge data can be created.

2.5. Analysis and Theory Formation

The most striking fact about the data is the enormous variation one finds through time and space. This variation contrasts sharply with the literature on funerary traditions in Taiwan, and, second, with statements that informants produce when explaining their traditions. The data thus not only question the foundations of established research on funerary rites, but also on research that uses informants as a source of information. In fact, the relations between publications, informants' opinions and a national ideology become palpable to such an extent that a scientific approach has to look into the involvement of ideologies in the transformation of cultural practices. The DH-approach is quite suitable for this endeavor: Digitizing scientific and political publications that stand in relation to funerary rites, we could contrast publications with the reality they pretend to describe and reveal their influences on social practices. More particularly, we could show how a governmental publication influenced the way that Taiwanese refer to their ancestral home through tombstone inscriptions .

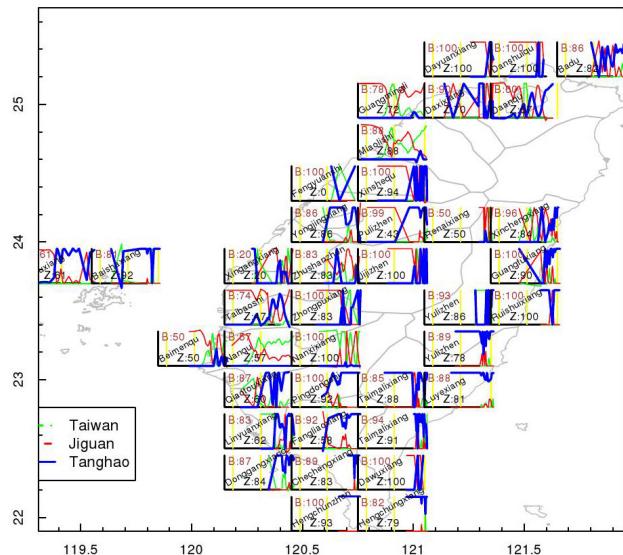


Fig. 3: An example for the variation through time and space. The onset of the place-name type 'tanghao' (blue) between 1900 and 2000. Adjacent regions may show similar or very different patterns in the development. B refers to the correlation to the Baijiaxing, a century-old book which

specifies which surname matches which tanghao. High correlations identify the place-name as a literary reference.

References

- thakbong.dyndns.tv
www.dans.knaw.nl
- Sperber, Dan** (1996). *La contagion des idées*. Paris: Odile Jacob.
- Bourdieu, Pierre** (1979). *La distinction: Critique sociale du jugement*. Collection Le sens Commun Paris: Éditions de Minuit.
- de Certeau, Michel and Giard, Luce and Mayol, Pierre** (1980/1990). *L'invention du quotidien: Arts de faire*. Paris: Gallimard.
- www.postgresql.org
- Holl, Stephan and Plum, Hans** (2009). *PostGIS*. GeoInformatics 3: 34–36.
- Chen, Shao-hsing and Fried, Morton** (1968). *The Distribution of Family Names in Taiwan: Volume I, The Data*. Taipei: National Taiwan University & Columbia University.
- Streiter, Oliver and Goudin, Yoann and Huang, Chun (Jimmy) and Lin, Ann Mei-fang** (2012). *Matching Digital Tombstone Documentation to Unearthed Census Data*. International Journal of Humanities and Arts Computing 6, 1-2: 57-70.
- Streiter, Oliver and Goudin Yoann** (2014). *The Tanghao on Taiwan's Tombstones: The Recuperation of Tactics for a National Space*. Archivi Orientalni.

Future Development of a System for Annotation and Linkage of Sources in Arts and Humanities

Subotic, Ivan

ivan.suboti@unibas.ch
 University Basel, Switzerland

Kilchenmann, André

a.kilchenmann@unibas.ch
 University Basel, Switzerland

Schweizer, Tobias

t.schweizer@unibas.ch
 University Basel, Switzerland

Rosenthaler, Lukas

lukas.rosenthaler@unibas.ch
 University Basel, Switzerland

1 Introduction

Since the late 90s, a large number of digitization projects relevant to research in the humanities have been carried out. The quality of the digital objects produced by these digitization campaigns most often meets the demands of a "digital facsimile". Since - with the exception of text files such as PDF, plain text, etc. where a full text search may be appropriate - digital objects are hardly searchable directly, associated metadata are needed to enable navigation within a collection of digital objects. It should be expected that this simplified accessibility and availability of digitized sources has fundamentally changed research in the humanities by allowing more efficient and broader research methods. However, it seems that this is not yet the case. The reason is that there are very few digital tools available to support the qualitative and comparative methods required in source based research in the humanities.

In the following, we will look at three use-cases which exemplify our vision for a research environment in the digital Humanities.

Digital Humanist. Susan, a digital humanist, is working with digitized manuscripts. For her research, she needs to transcribe, annotate, and link her annotations, regions of interest, and transcriptions with each other. By employing SALSAH as her work environment, Susan can work on the digitized manuscripts in a fully digital workflow.

Long-term Accessibility of Digital Research Data. Jim is at the stage of finishing up a five-year project, and needs to deposit his research results and the data accumulated during his research. The results and the digital data need to be still accessible in the long-term, even after the funding has long since ended. Jim can export his digital research data to an institution deploying SALSAH which will take care of their long-term accessibility.

Linked Open DataWorkbench. Karen's research is based on materials which are provided in different repositories around the internet. She wants to be able to combine, annotate and create links between those digital objects. Also she would like to share her results and allow other researchers to use them. By using SALSAH, Karen can connect to external resources shared over custom APIs or SPARQL endpoints, and work with the data as if it were stored locally. Using the SALSAH API and the provided SPARQL endpoint, other researchers can build upon her work.

SALSAH (System for Annotation and Linkage of Sources in Arts and Humanities) version 2.0 is currently under development at the Digital Humanities Lab (DHLab) of the University of Basel, and represents a browser based VRE that will respond to requirements described in the three scenarios above.

The main contribution of this paper lies in the description of novel approaches taken in the design of SALSAH 2.0, leading to new features and possibilities.

The remainder of this paper is organized as follows. In Section 2, we introduce newly developed features and Section 3 concludes.

2 SALSAH

SALSAH integrates digital (re)sources, metadata, research data, and relevant working tools. Using SALSAH, researchers are able to: (1) simultaneously visualize multiple digital objects (e.g., facsimiles, images, texts, transcripts, sound and video), (2) annotate digital objects and share these annotations with others (3) establish relations (links) between digital objects and annotate these relations, (4) access and integrate external data sources (e.g., digital libraries) so that the VRE tools may be applied to these sources without the need for local duplicates, and (5) transcribe manuscripts, speech and video.

2.1 Software Architecture

The software is based on a multi-tier architecture in which application logic is distributed between (1) a client application ("front end") which users interact with, (2) a more or less centralized server ("back end"), and (3) local and/or external data providers which provide the sources that users can work on. The SALSAH software architecture is depicted in Figure 1.

While SALSAH has the capability to function as a repository for digital sources, this is not its primary goal. There are many repositories of professionally digitized sources, and it makes no sense to duplicate their content in yet another repository. Following a logical separation of annotation tools and digital representations, SALSAH provides the basis for referencing sources without having to store them itself. Furthermore, SALSAH can provide annotation, linking information, and metadata to an external data provider via the SALSAH API (as long as the external request has access rights), as well as over a read-only SPARQL endpoint that provides LOD (Linked Open Data). We expect SALSAH in the long term to evolve into a true distributed P2P system.

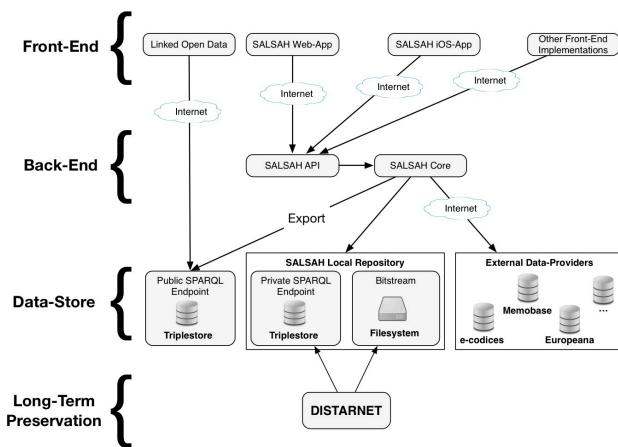


Fig. 1: Software Architecture of SALSAH

2.2 Data-Model

The data model is based on the Resource Description Framework (RDF), the Resource Description Framework Schema (RDFS), and the OWL 2 Web Ontology Language all proposed by the World Wide Web Consortium (W3C) for implementing the Semantic Web. This metadata model makes it possible to describe digital objects in a very flexible way, and to create links and relation between any objects (which are called "subjects" in RDF terminology). It is based on statements in the form of subject-predicate-object expressions about these digital subjects. Any number of such expressions can be used to describe subjects and their relations.

A given set of predicates is called a vocabulary, and can be used to implement standard metadata schemes such as Dublin Core. Within SALSAH, different vocabularies may be used at the same time to describe a given subject. Since the value of an RDF expression may itself be a subject, RDF allows for a network-like representation of knowledge about a subject and its relations to other subjects. This metadata model is subject-centric, in the sense that for each digital subject, an individual set of predicates may be assigned, in contrast to the relational data model, which is much more restrictive in its ability to assign data field to subjects. Hence, the data model used in SALSAH is especially well suited to the humanities, in which a flexible, qualitative coverage of metadata is essential. Figure 2 (a) depicts an excerpt from the SALSAH ontology, showing how a projects own metadata schema can be incorporated into SALSAH, and (b) a small part of the graph depicting an incunabula of Sebastian Brant with the title "Das Narrensch".

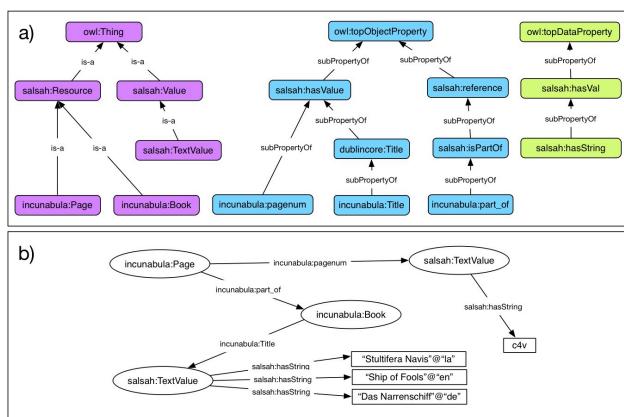


Fig. 2: An excerpt from (a) the SALSAH Ontology and (b) the incunabula of Sebastian Brant.

The data store consists of a native triple-store solution such as Jena, which serves the data over a SPARQL endpoint.

2.3 Versioning

SALSAH is a dynamic system in which data can be changed by users having the necessary access rights at any time. In order to use SALSAH as a citable repository, methods will be implemented to "freeze" a subset of the data and thus provide versioning. In order to solve this non-trivial problem, SALSAH will use the concept of temporal RDF, in which each element in the RDF graph of a certain granularity will be enriched with temporal information regarding its validity. For example, if the title of a book is changed, the old version is not overwritten, but is instead marked as valid up to the time when the change occurred, while the new title is marked as valid from then on. This allows users to retrieve the state of the RDF graph at any point in time.

Versioning will lead to the concept of a new form of electronic publication. While e-papers and e-journals basically mimic the behavior of their paper equivalents an annotated network of citable sources and links represents a novel form of publication. The reader will be able to navigate through the network and extract his or her own perspectives on the knowledge represented by the interconnected digital objects. This may be the first attempt, within academic publishing in the humanities, to go beyond the phenomenon in which "new media first mimic older media", as noted by Marshall McLuhan.

2.4 Digital Long-Term Preservation

DISTARNET (DISTRIBUTED ARchival NETwork) is a distributed, autonomous long-term digital preservation system. Essentially, DISTARNET exploits dedicated processes to ensure the integrity and consistency of data with a given replication degree. At the data level, DISTARNET supports complex data objects and the management of collections, annotations, and arbitrary links between digital objects. At process level, dynamic replication management, consistency checking, and automated recovery of archived digital objects is provided, using autonomic behavior governed by preservation policies without any centralized component

DISTARNET will be implemented as a layer underneath the SALSAH local repository, and provide long-term preservation of the digital objects and associated metadata.

3 Conclusion

While the change from the analog to the digital domain makes sources available on the desktops of scholars and researchers, a real paradigm shift in source-based research requires new tools. Virtual Research Environments such as SALSAH may provide the necessary tools to gain a novel, computer-aided knowledge representation that is well-suited to the needs of humanities research. These tools will undoubtedly change the way research is done in the humanities. They will help researchers organize and retrieve knowledge more efficiently, and may disclose hidden relationships between sources, among other things, but they will not replace the researchers' ingenuity and intuition. SALSAH is in use by several research projects within the University of Basel, and has sparked interest on an international scale.

References

- Rosenthaler, Lukas** **Virtual Research Environments** (2012). *A New Approach for Dealin with Digitized Sources in Research in Arts and Humanities* in: Claire Clivaz u.a. (editors): *Reading Tomorrow. From Ancient Manuscripts to the Digital Era*, Lausanne2012, S. 661-670, Ebook on <http://www.ppur.info/lire-demain.html>

Rosenthaler, Lukas (2012), *Schweizer, Tobias SALSAH - eine webbasierte Forschungsplattform für die Geisteswissenschaften*, in: *Bulletin der Schweizerischen Akademie der Geistes- und Sozialwissenschaften*, Bern

Rosenthaler, Lukas (2011) *Entwicklung einer Web 2.0-Applikation zur Präsentation un Erforschung der Basler Frühdrucke*, in: Karin Krause und Barbara Schellewald (editors), *Bild und Text im Mittelalter*, Böhlau Verlag Köln

- Schweizer, Tobias, Rosenthaler (2011), Lukas SALSAH - eine virtuelle Forschungsumgebung für die Geisteswissenschaften**, in: Dr. Andreas Bienert, Dr. Frank WeekendDr. James Hemsley, Prof. Vito Cappellini (editors), EVA 2011 Konferenzband pp. 147-153 GFai Berlin
- F. Manola and E. Miller**, "RDF Primer," tech. rep.
- D. Brickley and R. von Guha**, "RDF Vocabulary Description Language 1.0: RDF Schema," tech. rep.
- W3C OWL Working Group**, "OWL 2 Web Ontology Language," tech. rep.
- Dublin Core Metadata Initiative**. <http://dublincore.org/>.
- JENA**. <http://jena.apache.org/>.
- C. Ogbuji**, "SPARQL 1.1 Graph Store HTTP Protocol," W3C working draft, W3C, May 2011. <http://www.w3.org/TR/2011/WD-sparql11-http-rdf-update-20110512/>.
- C. Gutierrez, C. Hurtado, and A. Vaisman** (2005), "Temporal RDF," in European Semantic Web Conference The Semantic Web Research and Applications, vol. 3532/2005, pp. 93-107.
- J. Tappolet and A. Bernstein** (2009), "Applied temporal RDF : efficient temporal querying of RDF data with SPARQL," The Semantic Web: Research and Applications, no. June.
- McLuhan, E. and Zingrone, F.** (1995) (eds) *Essential McLuhan*. New York: BasicBooks
- I. Subotic**. *A Distributed Archival Network for Process-Oriented Autonomic*

A Large Database Approach to Cultural History

Sullivan, Brenton
brenton.sullivan@ubc.ca
 University of British Columbia

This panel introduces the work completed since 2012 by the Database of Religious History (DRH)¹. The DRH is one of the flagship initiatives of the newly established Cultural Evolution of Religion Research Consortium (CERC)², based at the University of British Columbia and Simon Fraser University in Vancouver but involving eight Partner Institutions and over 55 collaborators from all over the world. CERC is funded by a 6-year, \$3 million Partnership Grant from the Social Sciences and Humanities Research Council (SSHRC) of Canada³, along with approximately the same amount in matching funds from the host institutions, partner institutions, and other partner organizations. The DRH aims to bring together, in a standardized, systematic form, data on sociocultural history from across the world and throughout history, from the earliest archeological records up to approximately 1500-1600 CE.

The DRH was originally focused on the collection and analysis of data pertaining to religious traditions—hence, its name—but since the summer of 2013 it has also had as its mandate the collection of other historical variables for China and Mesoamerica. These variables pertain in particular to historical polities, social complexity, natural resources, and warfare. Beginning in 2014, the DRH has extended its collection of all types of historical data to the other selected world regions (see below). The DRH thereby provides a reliable system for statistically measuring social complexity of past human civilizations and for identifying the causes of such complexity.

The panel comprises four papers that provide a the disciplinary and historical context of the project, a frank assessment of the project's methodologies, a case study of religious and social history in Latium, and an overview of the suite of digital tools that have made the project possible. The ninety-minute period will be shared by five presenters (the technical paper has two co-authors), who will present in an organized yet dynamic fashion the history of the project since its conception in 2012. The paper by Brenton Sullivan (a post-doctoral research fellow at the University of British Columbia) focuses on the execution of the DRH, elaborating on the challenges involved with soliciting data from experts from around the globe and the most recent results of the project.

The paper by Fred Tappenden (post-doctoral research fellow at McGill University) presents the theoretical challenges involved in defining religious traditions for the sake of quantitative analysis. Next, Carson Logan (technical specialist for DRH at the University of British Columbia) and Michael Muthukrishna (technical specialist for DRH and doctoral student at British Columbia) provide a live demonstration of the Database of Religious History and visualizations of the results of the research while introducing the technical hurdles encountered in the process. Finally, the paper by Edward Slingerland, the director of the DRH and Professor at the University of British Columbia, places the DRH in its disciplinary and historical context and summarizes its conception, its longterm goals, the practical and theoretical challenges it has faced, and how it has already begun to influence various fields of the humanities, particularly the academic study of religion. All five presenters have enthusiastically agreed to present in Lausanne.

References

1. See www.religiondatabase.arts.ubc.ca . A collaboration between DRH and another research project based at the University of Oxford, the Seshat Global Databank, dissolved in February of this year. Although DRH's data and the arguments made in our panel are our own, we are very indebted to Seshat and its directors for much of the inspiration to commence a project of such magnitude.
2. See www.hecc.ubc.ca/cerc/project-summary .
3. See www.sshrc-crsh.gc.ca .

Digitizing Women's Literary History: The Possibility Of Collaborative Empowerment?

- Suzan, van Dijk**
 Huygens ING - Royal Dutch Academy of Arts and Sciences, Netherlands
- Dekker, Ronald**
 Huygens ING - Royal Dutch Academy of Arts and Sciences, Netherlands
- Partzsch, Henriette**
 St. Andrews University, UK
- Prats Lopez, Montserrat**
 Vrije Universiteit Amsterdam, Netherlands
- Sanz, Amelia**
 Complutense University Madrid, Spain
- Filarski, Gertjan**
 Huygens ING - Royal Dutch Academy of Arts and Sciences, Netherlands

Introduction

For several years, we have taken initiatives in view of clarifying women's realparticipation in the European literary field before the early 20th century – as opposed to the relative absence of women in literary histories of the 20th century. We have worked in this large field through a series of successive projects funded on a national as well as a European scale. At present, the HERA project Travelling TexTs 1790-1914. The Transnational Reception of Women's Writing at the Fringes of Europe (2013-2016) and the CLARIN-NL project Connections Between Women and Writings Within European Borders (COBWWWEB, 2013-2014) provide the context of our presentation. Roughly speaking: the HERA project is about the content, while in the CLARIN project developers are preparing a new structure, allowing the database WomenWriters to connect to other – either structured or editing – projects in the field of women's literature: for the sake of testing, in the first instance, to the Swedish Selma Lagerlöf Archive, the Norwegian Female

Robinsonades and the Serbian Knjizenstvo project; others will follow.

The intended large-scale approach is meant as a complement to ongoing individual digitizing projects, such as the Huygens ING “Digitizing Isabelle de Charrière’s letters” using the eLaborate edition project. Admittedly, “literary women” are benefitting from the new drive to digitize, but we are not yet able to feel the full benefit of these digitized texts. As Jacqueline Wernimont (2013) states, “simply saving women’s work in digital form” is not enough. The women’s texts are often presented in isolation from their historical reception context – which makes it impossible to evaluate their historical importance.

The large-scale and the individual

Referring to Labrosse 1985, we propose to take this reception context as a starting point for a large-scale approach on female authorship from previous, older periods in time. It represents the other end of the communication process in which these women engaged, and it helps us to select authors and texts that should be studied in more detail. When using these reception documents as an “entry” to the texts themselves, the emphasis obviously is on what struck the contemporary reader, in a positive and negative sense: we are thus in the middle of a dialogue.

Putting this data – with the appropriate metadata – in our online database (discussed and tested during an earlier COST financed phase of the collaboration) allows us to roughly situate these authors and works before analyzing them. Giving these women their own place in the virtual representation of the literary field, where communication, circulation and transmission can be made visible, provides context and promotes understanding on a larger scale.

Linking for understanding relevance

It is, indeed, not the sheer presence of these women’s texts on the Internet that advances scholarship. It is the possibility of understanding their relevance – which has often been systematically denied, without reference to empirical data by way of arguments. Understanding relevance cannot be reached by any unprepared reading of the writings. Approaching texts from a “prepared” perspective obviously requires time, but frees these texts from prejudices that are inherent in the late 19th and 20th-century literary historiography of women’s authorship and that – although denounced by Virginia Woolf in *A Room of one’s own* (1929) – inevitably influences even our post-feminist students.

On the practical level, the connection between the structured database content, on the one hand, and the online edited texts to be studied, on the other, can be made visible by using the same terminology for (1) distinguishing database categories, and (2) annotations in the digitized text in editing platforms such as eLaborate. Linking this data is particularly essential for research in women’s literary history because of the small amount of information available, and therefore the need to compare women authors, postulating that problems encountered by one author are experienced similarly by her colleagues.

Use cases

The objective of this presentation is to highlight the importance of ICT tools used in connection and in complement with each other, for research in domains (not just the one of women’s history) which have fallen behind. We will illustrate this using two examples, representing the two ends of literary communication:

1. A female sender, the Dutch-Swiss author writing in French Belle van Zuylen/Isabelle de Charrière (1740-1805), and her international (male + female) reception : she has her place in the WomenWriters database, and her letters are being digitized within the eLaborate project;

2. A male receiver, the Dutch 19th-century literary critic Conrad Busken Huet, who commented upon important numbers of (Dutch and foreign) women authors : he is present in the WomenWriters database, and his critiques are presented (without the possibility of annotation) in the Digital Library of Dutch Literature DBNL.

Briefly describing these two cases we want to illustrate the way in which this collective and transnational research in women’s literary history not only relies on combining several types of ICT tools, connecting different kinds of data (empirical data, “captia”, (references to the) primary and secondary texts), but also requires the participation of different categories of collaborators: not only professional researchers and ICT specialists but also volunteers.

This “mixed” collaboration is needed at a time when computers are still unable to communicate through the use of different languages. As this project collates women authors from all over Europe different tags and categories (denominated in English) are needed to refer to texts and aspects of texts in different languages. This can currently only be done by judgment, by readers who understand which elements of the (narrative) texts are characteristic or a-typical.

Many of the volunteers obviously will be women: the potential readers of both our research output and of the works written by “our” early women authors. They represent, typically, a category of people who will be “empowered” by the access that is given to their foremothers and potential role models....

Administrative obstacles

This complicated interconnection of collaborators may prove to be difficult. However, scholarship in the academy can no longer be fulfilled exclusively by the work of the professoriate. Technological innovations are not just a matter of devices and tools. They concern social practices; they concern users and uses. We can no longer ignore that, next to professionals (researchers, information technologists, librarians, and even students), there are “end users”, who are in fact essential members of our interdisciplinary community; they justify our very activities. It is important – given the precarious position of women in present-day society and the small number of historical role models – to “spread the word” about these early writing activities and the women behind them.

This massive, complex collaboration is not, at present, recognized by institutions and stakeholders; consequently, young and senior researchers, as much as students, are often not ready to invest themselves in this kind of collaborative project. In our presentation, we will also denounce the gaps between digital technologies, the willingness of potential volunteers and the entire educational and academic system, and we will propose some possible strategies.

References

Bergenmar, J., Olsson, L., (2012). *Connecting European Women Writers*. The Selma Lagerlöf Archive and Women Writers Database, in Digital Humanities 2012: Conference Abstracts. Hamburg: Hamburg University Press, pp. 113-114. www.dh2012.uni-hamburg.de/wp-content/uploads/2012/07/HamburgUP_dh2012_BoA.pdf

Bergenmar, J., Marzec, L., Sanz, A., (2013). *Learning to hack the Literary History – teaching Transnational Women’s Writing Digitally*, contribution to final conference of COST Action IS0901 “Women Writers In History”: Female authorship in Europe: Networks and obstacles. The Hague, June 2013. www.womenwriters.nl/index.php/Learning_to_hack_the_Literary_History

Jeay, J., Sinclair, S., (2012). *En quête d’amitié. Approches méthodologiques pour l’analyse automatisée d’un corpus électronique*, contribution to SATOR conference Topiques de l’amitié, Victoria June 2012. web.uvic.ca/~amitie/index%20resumes.html

Labrosse, C., (1985). *Fonctions culturelles du périodique littéraire*, in Claude Labrosse, Pierre Rétat, L’instrument

- périodique, *La fonction de la presse au XVIIIe siècle*. Lyon: Presses Universitaires de Lyon, pp. 11-136.
- Van Dijk, S.,** (2012). *La correspondance d'Isabelle de Charrière en ligne*, Cahiers de l'Association Internationale des Etudes Françaises, 64, pp. 29-40. www.womenwriters.nl/images/8/81/2012_SvanDijk_01_lacorrespondance.pdf
- Van Dijk, S.,** (2012). *Amitié, solidarité et entraide féminines : Spécificités d'auteurs femmes ?*, contribution to SATOR conference Topicues de l'amitié, Victoria June 2012. web.uvic.ca/~amitie/index%20resumes.html
- Van Dijk, S., Hoogenboom, H., Sanz, A., Bergenmar, J., Olsson, L.,** (2012). *Data sharing, virtual collaboration, and textual analysis: Working on "Women Writers In History"*, in Digital Humanities 2012: Conference Abstracts. Hamburg: Hamburg University Press, pp. 527-529. www.dh2012.uni-hamburg.de/wp-content/uploads/2012/07/HamburgUP_dh2012_BoA.pdf
- Van Dijk, S., Prats Lopez, M.** (2013). *Participation as a way of creating new audiences ?*, contribution to final conference of COST Action IS0901 "Women Writers In History": Female authorship in Europe: Networks and obstacles. The Hague, June 2013. www.womenwriters.nl/index.php/Participation_as_a_way_of_creating_new_audiences_%3F
- Wernimont, J.,** (2013). *Whence Feminism? Assessing Feminist Interventions in Digital Literary Archives*, in Digital Humanities Quarterly 7/1. www.digitalhumanities.org/dhq/vol/7/1/000156/000156.html

Analysis of perspectives in contemporary Japanese novels using computational stylistic methods

Suzuki, Takafumi
Toyo University, Japan

Yamashita, Natsumi
Toyo University, Japan

1. Introduction

Perspective in novels has been an important subject of research in literary studies. Ishimaru (1985) defined perspectives as the viewpoint of narrators; she roughly classified perspectives in novels as the first-person perspective, where the central character narrates the story from his/her perspective, and the third-person perspective, where the omniscient narrator recounts the story from a neutral perspective. This is a basic classification of perspective in literature. These perspectives represent the spirit of the age, typically shown in the positivism in 19th century French novels (Ishimaru, 1985), and also affect a readers' impression of the characters and involvement in the work, and thus perspective is an important subject in literary studies.

Computational stylistics has been one of the important subfields of Digital Humanities. Using computational methods with digitized text materials, we can obtain systematic findings that can complement traditional qualitative analyses. Although computational methods can be powerful tools for investigating issues in literary studies, perspective in novels has rarely been analyzed with such method.

Against this background, we used computational stylistic methods, i.e., text classification and feature analyses by random forests machine learning methods, to tackle the perspective issue in literary studies. We selected Kotaro Isaka, who is a popular Japanese novelist, as the object of study; he explicitly switches perspective in his novels section by section, and this is an important reason for the popularity of his novels. Note that Haruki Murakami, another popular novelist, uses this perspective switching between two perspectives (Kudo

et al., 2012). However, Isaka uses more varied perspective-switching patterns (Yamashita and Suzuki, 2013). First, we generated text files and applied morphological analysis. We then conducted random forests text classification and feature extraction experiments using text-feature matrices for two of Isaka's novels. Then, we investigated (a) whether textual differences among perspectives can be detected or not, and (b) if detected, what types of textual characteristics contribute to the detection of perspective. By tackling these points, we will show the effectiveness of computational methods for analyzing the perspective issue in literary studies.

2. Data and methods

We selected the following novels by Kotaro Isaka, "Odyubon no Inori" (Audubon's Prayer; ADP, original 2000, pocket edition 2003) and "Gurasuhoppa" (Grasshopper; GHP, original 2004, pocket edition 2007) as objects. ADP is a work representative of the earlier period of the author's bibliography, and GHP is representative of the author's middle period. We used the pocket editions of these two novels because Isaka is known to revise manuscripts when his work is published in pocket editions. We constructed the texts using a OCR document scanner and manually corrected OCR errors. We also removed the rubi, i.e., kanas printed alongside kanjis. We applied morphological analysis using MeCab,[1] Japanese morphological analyzer.

We divided the texts into sections and assigned perspective tags according to the perspective signs assigned by the author. Regarding ADP, we united all character perspectives except Ito, the central character, because the number of perspectives for each character is small. Without unification of perspective, it was difficult to perform meaningful classification and feature analysis experiments. Thus, we used two tags, Ito's perspective and other characters' perspectives. The numbers of sections was 56 for Ito and 22 for other characters. It should be noted that Ito's section appeared after another of his section. Regarding GHP, we used three tags for the three main characters' perspectives (Suzuki, Kujira, and Semi) according to the signs assigned by the author. The sequences of these three characters' perspectives are essentially fixed, Suzuki first, Kujira second, and Semi third. In addition, the death of a character leads to the removal of that character's perspective. The numbers of sections was 17 for Suzuki, 15 for Kujira, and 10 for Semi.

We calculated the frequencies of morphemes and basic textual statistics, and then we constructed the text-feature matrices using the relative frequencies of morphemes appearing in each text. We applied random forests machine learning methods proposed by Breiman (2001) with these matrices as data and perspectives as labels. We calculated the valuable importance provided by random forests and extracted important variables for classification, which are effective for differentiation among perspectives. We selected the random forests method because it has shown the best possible performance for authorship attribution in Japanese (Jin and Murakami, 2007) and is effective for extracting and analyzing the features that contribute to classification in related tasks such as computational sociolinguistics (Suzuki, 2009).

3. Results and discussion

3.1. Basic observation

Table1. Basic data (ADP)

		Number of tokens			
	Number of texts	sum	mean	s.d.	c.v.
Ito	56	118042	2107.89	1712.37	0.81
Others	22	23290	1058.64	1315.31	1.24

Table 1 shows the basic data for ADP, the number of texts, and the sum, mean, standard deviation (s.d.), and coefficient of variations (c.v.) of the number of tokens for each perspective. It can be seen that Ito has more than 70% of all sections, and others have larger variances of the c.v. It is assumed that the larger variances were caused by the unification of characters.

Table 2. Basic data (GHP)

		Number of tokens				
	Number of texts	sum	mean	s.d.	c.v.	
Suzuki	17	51229	3013.47	1930.56	0.65	
Kujira	15	33453	2230.2	1251.70	0.56	
Semi	10	27153	2715.3	946.63	0.35	

Table 2 lists the basic data for GHP, the number of texts, and the sum, mean, s.d., and c.v. of the number of tokens for each perspective. The table shows that Suzuki has the largest section numbers and has the largest c.v. It is assumed that Suzuki's perspective includes both small and long sections.

3.2. Classification by random forests

Table 3. Classification results (ADP)

	Ito	others	error rates
Ito	55	1	0.02
Others	17	5	0.77

Table 3 shows the classification results obtained by random forests for ADP. Each column represents the original tags, and each row represents the results. It can be seen that 55 of 56 Ito texts were classified as Ito's. It is assumed that Ito's perspectives have special characteristics. In comparison, only 5 of 22 texts by others were classified as others and 17 of 22 texts by others were classified as Ito. It is assumed that these results were partly caused by the limits of our experiments; the number of Ito texts was much larger than others, and the text from several characters was merged.

Table 4. Classification results (GHP)

	Suzuki	Kujira	Semi	error rates
Suzuki	17	0	0	0
Kujira	0	15	0	0
Semi	0	5	5	0.45

Table 4 shows the classification results obtained by random forests for GHP. Each column represents the original tags, and each row shows the results given by random forests. It can be seen that all Suzuki texts were 17 were classified as Suzuki's, and all Kujira texts were classified as Kujira's. Only 5 of 10 texts were classified as Semi, and 5 other texts were classified as Kujira. It is assumed that there were special characteristics for Suzuki and Kujira's perspective; however, in comparison, Semi's perspectives were rather characterless and closer in nature to Kujira's texts. It is worth noting that both Semi and Kujira are assassins, and Suzuki is an employee; therefore, it is assumed that the fact that Semi and Kujira are similar characteristics indicates the author's intent to differentiate these two characters and Suzuki.

3.3. Feature analysis

Table 5. Top 20 important features (ADP)

	feature	readings	translation	pos	variable importance
1	僕	Boku	I	noun (pronoun)	0.01911
2	だ	da	be	auxiliary verb	0.00661
3	日比野	Hibino	Hibino	noun (proper)	0.00404
4	ん	n	-	noun, auxiliary verb, particles	0.00293
5	。	-	-	symbol	0.00267
6	を	wo	-	particle	0.00262
7	静香	Shizuka	Shizuka	noun (proper)	0.00253
8	」	-	-	sign	0.00246
9	声	Koe	Voice	noun	0.00214
10	しれ	shire	-	verb	0.00177
11	よ	yo	-	particle	0.00172
12	伊藤	Ito	Ito	noun (proper)	0.00165
13	かも	kamo	May	particle	0.00142
14	歯	Ha	Dent	noun	0.00126
15	に	ni	-	particle	0.00113
16	いや	Iya	-	exclamation	0.00106
17	島	Shima	Island	noun	0.00095
18	返事	Henji	Reply	noun	0.00094
19	目	Me	Eye	noun	0.00093
20	?	-	-	symbol	0.00092

Table 5 shows the top 20 variables that contributed to classification of ADP with English translations, indicates parts of speech, and shows the variable importance obtained by random forests. The variables include many proper nouns and content words such as "島" (Shima; Island) which simply represent contextual difference in the narrative. Table 5 also includes stylistic characteristics such as pronouns that represent the differences between the perspectives of Ito and others.

Table 6. Top 20 important features (GHP)

	feature	reading	translation	pos	variable importance
1	鈴木	Suzuki	Suzuki	noun (proper)	0.00947
2	妻	Tsuma	wife	noun	0.00938
3	比	Hi	-	noun (proper)	0.00812
4	亡き	Naki	dead	adnominal	0.00781
5	鯨	Kujira	Kujira	noun (proper)	0.00764
6	亡靈	Borei	ghost	noun	0.00699
7	僕	Boku	I	noun (pronoun)	0.00664
8	子	Ko	Ko	noun (proper)	0.00560
9	槿	Asagao	Asagao	noun (proper)	0.00524
10	西	Nishi		noun (proper)	0.00477
11	岩	Iwa	-	noun (proper)	0.00475
12	与	Yo		noun (proper)	0.00452
13	彼女	Kanojo	she	noun (pronoun)	0.00393
14	ねえ	nee	-	noun	0.00367
15	おまえ	Omae	you	noun (pronoun)	0.00354
16	長男	Chonan	eldest son	noun	0.00322
17	君	Kimi	you	noun (pronoun)	0.00297
18	つう	Tsuu	-	auxiliary verb	0.00268
19	なかっ	nakatt	-	auxiliary verb	0.00254
20	だろ	daro	-	auxiliary verb	0.00224

Table 6 shows the top 20 variables that contributed to the classification of GHP with translations in English, indicates part of speech and presents the variable importance obtained by random forests. Table 6 includes many [part of] proper nouns, indicating that they are the most important characteristics for discriminating the perspectives of the three main characters. In addition, Table 6 includes “つう” (Tsuu) and “ねえ” (Nee), which are style markers specific to several characters (e.g., Kujira) This indicates that these special style markers are also important characteristics for discriminating the perspectives among the three main characters.

4. Conclusion

This study analyzed the textual difference among perspectives in two contemporary Japanese novels. The results indicate that (a) respective perspectives have their specific textual characteristics, (b1) textual characteristics such as proper nouns that represent respective scenes are important for discriminating perspectives, and (b2) stylistic characteristics such as pronouns and nouns that represent styles of speech are also important. We conclude that computational stylistic methods can differentiate among perspectives in contemporary novels.

This study is a preliminary analysis of the study of perspectives using computational stylistic methods and is also part of an ongoing study of Kotaro Isaka's work. In future, we would like to further investigate the effectiveness of

computational methods for perspective issues and continue to analyze other work by Kotaro Isaka.

Acknowledgements

This study was supported by Grant-in-Aid for Scientific Research 23700288 for Young Scientists (B), from the Ministry of Education, Culture, Sports, Science and Technology, Japan. An earlier version of this study was presented at the 19th Annual Meeting of Japanese Natural Language Processing (NLP2008) at Nagoya University. This research includes revised and expanded content based on the gradation thesis presented by Natsumi Yamashita to the Faculty of Sociology, Toyo University.

References

- Breiman L.** (2001) *Random forests, Machine Learning*, Vol.45, pp.5-23.
- Isaka, K.** (2003) *Odyubon no Inori*, Sincho Bunko, Tokyo.
- Isaka, K.** (2007) *Gurasuhoppa*, Kadokawa Bunko, Tokyo.
- Ishimaru, A.** (1985) *Bunsyo ni okeru shiten*, Nihongogaku, 4(12), 22-31.
- Jin, M. and M. Murakami** (2007) *Authorship identification using random forests*, Proceedings of the Institute of Statistical Mathematics, 55(2), 255-268.
- Kudo, A., Murai, H. and A. Tokosumi** (2012) *Kyotsu go no fuchi to henka ni motoduku heiko keisiki syosetsu no monogatari kouzo*, Journal of Japan Society of Information and Knowledge, 22(3) 187-202.
- Suzuki, T.** (2009) *Extracting speaker-specific functional expressions from political speeches using random forests in order to investigate speakers' political styles*, Journal of the American Society for Information Science and Technology, 60(8), 1596-1606.
- Yamashita, N. and T. Suzuki** (2013) *Keiryo tekisuto bunseki wo mochiita syosetsu no shiten kenkyu: Isaka Kotaro wo rei to shite*, Proceedings of 19th Annual Meeting of the Association of Natural Language Processing (NLP2013), P1-3 (www.anlp.jp/proceedings/annual_meeting/2013.html). mecab.sourceforge.net

Integrating Score and Sound: “Augmented Notes” and the Advent of Interdisciplinary Publishing Frameworks

Swafford, Joanna

jes8zv@virginia.edu
University of Virginia

1. Introduction

While sound studies is experiencing a resurgence in literary criticism, academic arguments involving sound are nearly impossible to make in traditional print media. An article could include an excerpt of a score, but only scholars who can read music would be able to understand it. Likewise, articles or books could include audio files externally, as did Nicholas Temperley's special edition of Victorian Studies, which included a cassette tape with the songs discussed in the articles.¹ However, these solutions do not address the central problem: readers will have difficulty finding the exact musical phrases mentioned in articles, and those with less musical expertise will be left out of the conversation entirely. Newer options for incorporating music in academic articles include SoundCite (soundcite.knightlab.com), a tool that lets users embed sound clips in websites, Scalar (scalar.usc.edu/scalar), a publishing

framework that lets users annotate media, and the strategy of assigning a QR code to each audio excerpt and inserting these into a print article, as Jennifer Wood suggests.² None of these options integrates the audio with the score: SoundCite will only let users hear the audio, Scalar only supports textual annotations of media files, and QR codes require readers to have smart phones, which vastly limits the audience for the article. To address these problems, I have built two tools: "Songs of the Victorians" (www.songsofthevictorians.com), an archive and analysis of musical settings of Victorian poems with an interactive framework that highlights each measure of a score in time with its music, and "Augmented Notes" (www.augmentednotes.com), a public humanities tool that allows users who do not know how to program to build their own sites like "Songs of the Victorians."

2. Overview of "Songs of the Victorians"

"Songs of the Victorians" melds the archive and the scholarly article. It examines both high- and low-brow Victorian settings of contemporaneous poetry by integrating scores, audio files, and scholarly analytical commentary in an interactive environment to help users understand both the literary and musical elements of the argument. As an archive, it provides audio files of each song and archival-quality scans of first-edition printings of each score. For every song, the user can listen to the audio while each measure of the score is highlighted in time with the music, as the archive page for William Balfe's "Come into the Garden Maud" demonstrates (www.songsofthevictorians.com/balf/archive.html). The project also functions as a collection of scholarly articles in which each song includes an analysis of the song's interpretation of the text. When the commentary discusses a particular measure, the users can click on an icon of a speaker, which will play the relevant excerpt of the audio file and highlight the score so they can hear for themselves the effect the commentary describes, as in the analysis page for Caroline Norton's "Juanita" (www.songsofthevictorians.com/norton/analysis.html).

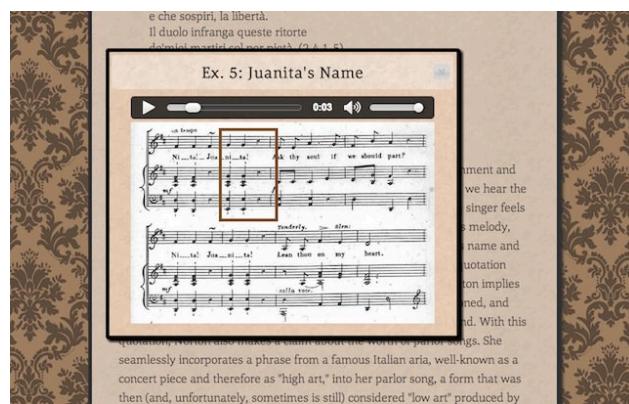


Fig. 1: Musical excerpt from the analysis page of Caroline Norton's "Juanita"

"Songs of the Victorians" includes Caroline Norton's "Juanita," Sir Arthur Sullivan's setting of "The Lost Chord," and two settings of Tennyson's *Maud*: a parlor song by Michael William Balfe and an art song by Sir Arthur Somervell. The site furthers scholarship for bibliographers, musicologists, Victorianists, and cultural studies scholars alike. More generally, this new framework, which enables critics to describe musical arguments to non-musicians, facilitates this interdisciplinary approach of bringing music and literature together. It also preserves the musical and cultural afterlives of well-known poems, as many of these scores have either disintegrated and been lost to time or are only available in select libraries. "Songs of the Victorians" empowers users regardless of musical training: those who cannot read music can overcome their feelings of intimidation at a musical score and can better understand the ideas described in the analysis,

whereas those who can read music will still benefit, since few people can hear in their mind the music on the page.

3. Overview of "Augmented Notes"

After the success of "Songs of the Victorians," I used its framework to produce "Augmented Notes" (<http://www.augmentednotes.com>), a generalized, public humanities tool to allow anyone to develop similar websites. "Augmented Notes" eliminates the need for users to understand programming by creating archive pages, like those from "Songs of the Victorians," which users can tweak and redesign. It is simple to use, as the site only requires audio files and images of the score to produce an archive page. After the audio and image files are uploaded, users are taken to a page where they click and drag to draw boxes around each measure (they can also edit the sizes and order of these boxes), indicating what portion of the score should be highlighted when that measure plays.

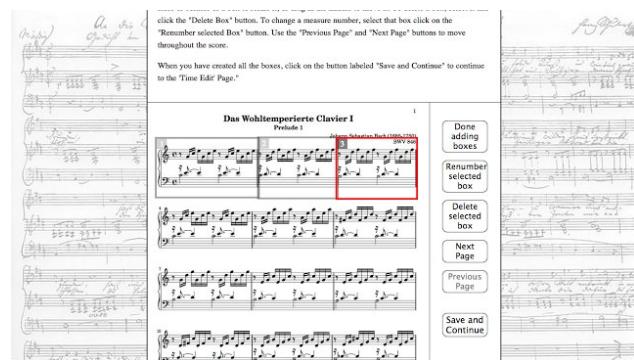


Fig. 2: Box-drawing page of "Augmented Notes"

Users can also optionally upload an MEI file--the TEI-based scholarly standard for music--for the score if they already have measure positions recorded in MEI. Users then set the times at which the highlighting box changes position through a "time editing" page.

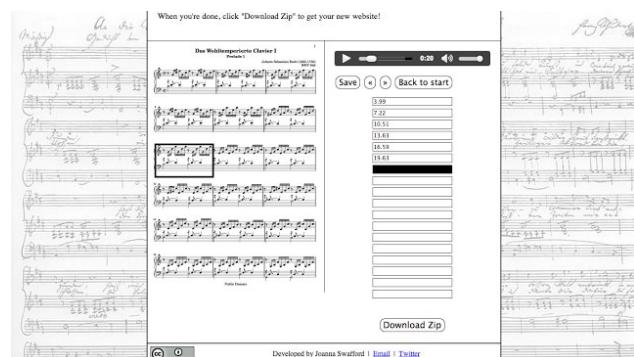


Fig. 3: Time editing page of "Augmented Notes"

The site brings together the measure and time information, saving them in a JSON file, which enables each measure of the song to be highlighted in time with the music. Users then click "Download Zip" to download a zip file with the HTML, CSS, and JavaScript files necessary for a complete archive page, which they can then restyle themselves.

"Augmented Notes" also has a sandbox (www.augmentednotes.com/example) through which users who would like to experiment with the technology but do not themselves have the requisite files can try it out. "Augmented Notes" is already being used by scholars, both for archival purposes (such as the "Performing Romantic Lyrics" project from the University of South Carolina) and for pedagogical purposes such as generating interactive scores for use in music classrooms. Since this tool produces websites with integrated audio and scores, it empowers users to preserve cultural archives, whether their materials include classical

music, unpublished manuscripts, popular music, or folk music and traditional tunes from around the globe.

4. Implications

This presentation will discuss the projects in greater detail, complete with live demonstrations and an explanation of the underlying technology to show their digital as well as scholarly innovations. I will explain the rationale for my choice of poems, settings, sound files, and editions for "Songs of the Victorians," as well as my plans for future collaboration and expansion for both projects. The presentation will illustrate the sorts of arguments that this framework can enable: for example, my examination of Sir Arthur Sullivan's setting of Adelaide Procter's "A Lost Chord" challenges the received interpretation that the poem merely describes a domestic, uncomplicated religious moment of transcendence. Likewise, Caroline Norton's "Juanita" has been considered a conventional song that preserves the traditional rules of courtship and parlor propriety, but my analysis of the music helps us see that it critiques the Victorian institution of marriage as imprisoning. I will conclude by exploring the ways in which "Songs of the Victorians" is itself a Victorian endeavor, as it uses new technology to collect, analyze, and bring together Victorian music and poetry, thereby giving voice to the silent page.

5. Funding

This project was made possible by fellowships from NINES and the Scholars' Lab.

References

1. Temperley, Nicholas (1986). *Music in Victorian Society and Culture: A Special Issue of Victorian Studies*. 30.1.
2. Wood, J. (2013). *Noisy Texts: How to Embed Soundbytes in Your Writing*. Burnable Books. Ed. Bruce Holsinger. burnablebooks.com/noisy-texts-how-to-embed-soundbytes-in-your-writing-a-guest-post/ (Accessed 30 October 2013).

Realizing the democratic potential of online sources in the classroom

Sweeny, Robert C.H.

rsweeny@mun.ca

Memorial University of Newfoundland, Canada

Burton, Valerie C.

Memorial University of Newfoundland, Canada

1.1 Overview

How might students understand a digitised edition of a major documentary series, created at the height of empire and that is perhaps the largest repository about working people from the mid-19th to the mid-20th centuries? Would differing forms of digitisation affect their understandings of the series? Our research results indicate poorly in answer to the first question and, surprisingly, not at all on the second, because understanding proved to be less a question of form than content. To realise the potential created by the radically democratised access that online sources permit, we first need to understand and respect how 21st century undergraduates see their world. Rather than focusing on form, we need to explain how their experiences as youth in a neo-liberal world on the brink of ecological disaster relate to questions of power, hierarchy and resistance in the past.

1.2 Methodology

In the preliminary stage individual qualitative and intensive examinations of undergraduate students were carried out using two fundamentally different editions of the same digitalised document. One half used a traditional textually-introduced document, the other accessed the same information through individual optional Google pins distributed throughout the document. This research was then supplemented by in-class experiments and work with graduate students and colleagues.

2.1 Preliminary questions

The dramatic increase in primary source material available online fundamentally transforms teaching and research in the humanities. Material that was until recently only available to hundreds or at best thousands of people at select, often unique, repositories is now virtually available to potentially millions of people. Historical sources are an important part of this completely unprecedented and radical reorientation of archival practice. The sheer scale of this newly democratised access to the sources of humanity's story is already extraordinary, and as more and more of the archival treasures long-housed in former imperial centres becomes available for research, and importantly for our purposes teaching in classrooms around the world, the potential for challenging Eurocentric conceptions of our past is great indeed.

This potential will, however, only be realized if our students are equipped with the tools they need to critically analyse these online resources. But is the historical literacy needed to understand a manuscript the same when encountered as a virtual source? Can a student simply transfer the knowledge and skills learnt in supervised archival research to working on the web? Are there different ways of knowing, distinct epistemologies, which are more appropriate to the qualitatively unique ontology of online sources? If so, should we be using the conventions of the digital world to navigate through these virtual sources? (Krug, 2006) How might such techniques affect the necessary respect of the historical distance between a conscientious researcher in the present and the source from the past that she or he is examining?

When you hold a centuries old artefact in your hands, feel its weight, hear the paper crinkle, notice a stain or perhaps just react to the dust, you sense a connection to the past that is simultaneously humbling and enriching. This experience has been at the heart of research in the humanities for centuries. (A. Burton, 2005 & Steedman, 2005) Can a virtual encounter be as meaningful? (Hayles, 2001 & JDH, 2012) Indeed how useful is to think of these virtual representations as being from the past? Are they for our students anything more than a brief illusory encounter with a largely incomprehensible past?

Discussion of these wide-ranging and difficult questions, posed as pedagogues of history, led the authors of this paper to engage in an on-going exploration we call "Explaining ourselves." An urgency fueled our discussions, as increasingly we realized that our undergraduate students see the world in fundamentally different ways than those we taught only a short time ago. So how might they encounter a major documentary series that was created at the height of empire, and is perhaps the largest repository with documentary consistency about working people from the mid-19th to the mid-20th centuries? Close to hand we had the basis for an answer.

2.2 Finding the answers

Memorial University of Newfoundland is a public university with little in the way of endowments. It is the only university in a province that has the highest poverty rates in Canada. Until recently, MUN has been a place where undergraduate teaching was respected. In the early 1970s, young historians raised the necessary funds from the federal government and convinced the local university administration to acquire the bulk of a collection documenting merchant seafarers of the British Empire, held by the Public Record Office in London. (Matthews, 1974) Covering more than a century, the Crew Agreements

maintained by MUN's Maritime History Archive document the workforce of three-quarters of all ocean-going vessels within the British Empire between 1863 and 1939, with declining numbers continuing until the early 1970s. It provides detailed information about the tens of millions of men and women who served aboard the largest merchant marine in history.

Completion of the major public history initiative *More Than a List of Crew* (V. Burton, 2011) provided the impetus for our study. By asking how would students read digital editions of these complex, multi-layered documents, she reinvigorated her co-author's decades-long engagement with the use of computers in the classroom (Sweeny, 1988-2010).

We developed two digital editions of the same crew agreement. The first had an extensive, multiple-screen-length, textual introduction, analogous to an introduction to a historical document in a scholarly edition in print. The second used Google pins to provide location specific information to help the user navigate the document. We created two sets of colour-coded pins. The first reproduced in bite-size portions the contents of the textual introduction. The second drew attention to any references to a particular individual onboard. A crew agreement from another vessel a decade later involving the same seafarer was also made available for the students to examine.

We worked with fourteen undergraduate Arts and Science students in individual sessions. We recorded the screen actions for each session. Working alone, students were given up to two hours to familiarize themselves with the document. Half worked with the text edition and half worked with the annotated one. Then the students had two hours to complete two exercises. The first was content-oriented and used a multiple-choice/short answer quiz. The second involved the second crew agreement mentioned above, which was not annotated, but the students could consult the earlier documentation. They were asked to write a brief analytical essay engaging both primary documents. These written assignments were followed up by focused conversations where we asked the students to explain their answers and to identify the problems they had encountered in working with the documents. As we progressed this de-briefing became more effective as we realized the merit of asking students how they would explain these documents to a friend.

In a second phase of the research we broadened our pool. We asked select graduate students in history and colleagues from our department to participate. Students in a third year historical methods class as well as students in a first year introductory course spent a 50 minute class with the documents after having visited the archives, formally in the first case and virtually in the second. They were then asked for written feedback.

The results indicated no significant differences in understanding of the documents, both groups fared poorly. Furthermore, students with substantial historical training did not show appreciable differences. Thus, our presumption that form mattered was misguided. Instead, student comments and their observed navigation practices both strongly suggest their need to engage directly with the documents in ways which use their existing understandings if they are to explore new ways of seeing.

2.3 Where from here?

We need to consciously transcend the form/content divide, if we are to engender in our students both an appreciation of the internal logic of a source and the ability to engage concept and evidence. In this particular case, how a crew agreement embodies the unequal power relationships between seafarers, masters, ship owners and the state would be the most fruitful pedagogical focus. What are they agreeing to and why? Crew agreements, like almost all historical documents, record unequal relations. Young people today live in an increasingly unequal world and this is the key to a progressive pedagogy that opens up our troubled present to our many and varied pasts. Used critically, it might yet allow our students to realize the democratic potential of all those newly accessible documents.

References

- Anderson, Steve F.** (2011). *Technologies of History: Visual Media and the Eccentricity of the Past*. Hanover, N.H.: Dartmouth College Press.
- Berger, John.** (2007). *Hold Everything Dear: Dispatches on survival and resistance*. New York: Pantheon Books.
- Booth, A.** (2006). "Perspectives on the research-teaching relationship in history." www.hca.heacademy.ac.uk/assets/hca/documents/case_Studies/snash/booth.doc
- Burton, A.** (2005). "Introduction: Archive Fever, Archive Stories". In *Archive Stories, Facts, Fictions and the Writing of History*. Durham & London, Duke University Press: 1-24.
- Burton, V.** (2011). *More Than a List of Crew*, www.mun.ca/mha/mlc
- Hayles, N. K.** (2001). "The Transformation of Narrative and the Materiality of Hypertext." *Narrative* 9, 21-39.
- JDH, the editors** (2012). "The Difference the Digital Makes." *Journal of Digital Humanities*, 2, October.
- Krug, Steve.** (2006). *Don't Make Me Think: A Common Sense Approach to Web Usability*. Berkley: New Riders Publishin
- MacDonald, T** (1996). Ed. *The Historic Turn in the Human Sciences*. Ann Arbor: University of Michigan Press.
- Jimerson, R. C.** (2006). "Embracing the Power of Archives." *The American Archivist*, 69,1: 19-23.
- Matthews, K.** (1974). "Crew Lists, Agreements, and Official Logs of the British Empire, 1863-1913." *Business History*, XVI, 1: 78-80.
- Schmidt, H. C.** (2012) "Media, Millennials, and the Academy: Understanding the State of Media Literacy within Higher Education." *Journal on Excellence in College Teaching* 23.4 (2012): 53-75.
- Steedman, C.** (2002) *Dust: the archive and cultural history*. New Brunswick: Rutgers University Press.
- Sweeny, Robert C.H** (1988) *Les relations ville/champagne, le cas de bois de chauffage*. Montréal: Éditions du MBHP.
- Sweeny, Robert C.H.** (1997/8). "The Past in the Present: Part I Epistemological challenges of computer-assisted teaching, Part II, Methodological reflections on computer-assisted teaching." www.chashcaccommittees-comitesa.ca/cchccchi/Doc/CHA98_Sweeny.htm
- Sweeny, Robert C.H.** (2010) *Montréal, l'avenir du passé: le 19ième siècle*. St John's: MMS Atlantic.

Stylometry of Collaborations: Dickens, Collins and their collaborative writings

Tabata, Tomoji

tabata@lang.osaka-u.ac.jp
GSLC, University of Osaka

1. Introduction

The Victorian author Charles Dickens was among the first publishing entrepreneurs to run mass-produced weekly/monthly magazines on a successful commercial basis. He employed many 'salaried staff writers' (Nayder, 2002), who had to write under anonymity, including Elizabeth Gaskell, Adelaide Anne Proctor et al., in *Household Words* and *All the Year Round*, the journals 'conducted by' Dickens (Stone, 1968; Thomas, 1982; Allingham, 2011).

On the other hand, Dickens collaborated with his younger contemporary Wilkie Collins on a number of stories, typically for the Christmas Numbers of his journals. While some of their collaborative pieces were written with the assistance of other staff writers, four works are known to have been co-authored by Dickens and Collins alone (Nayder, 2002): *The Frozen Deep* (1857), *The Lazy Tour of Two Idle Apprentices*

(1857), *The Perils of Certain English Prisoners'* (1857), and *No Thoroughfare* (1867). The four collaborations can be seen as betokening what appears to be a firm presence of Collins, a foothold he had gained, in the Dickens circle by the time he and Dickens launched into the joint works beginning in 1857.

These collaborative writings vary in design and style from one another as well as in theme and setting. In some cases, one chapter can be read as radically different from another due in large to the varying proportion of contribution by each of the duo: some chapters were written either Dickens or Collins alone, while there are other chapters that were jointly written although the extent of collaboration has yet to be identified quantitatively.

In order to provide new insight to the nature of collaborative authorship, the present study applies a series of stylometric techniques: (1) Craig's extension of Burrows's *Zeta* test for reliably extracting author markers from a large number of candidate words; (2) Cluster analysis based on Burrows's *Delta* distance measure (Burrows, 2002) to compare the collaborations with the canonical works of Dickens and Collins; and (3) Rolling *Delta* (Eder, Kestement and Rybicki, 2013) in an effort to detect authorial takeovers or to estimate the extent of contribution by each of the two authors in their collaborative writings.

2. Single authorship and mixed authorship

Although the lack of byline makes it difficult to determine the authorship of Christmas numbers for Dickens's periodicals, the account book in the office of Household Words helps identify many of his collaborators (Thomas, 1982). Table 1 shows bibliographic details for the four collaborative works between Dickens and Collins.

The Frozen Deep was originally written as a drama in three acts. Collins drafted the manuscript and Dickens heavily revised it. The script of the drama remained unpublished until 1866, when Collins altered it single-handedly, getting rid of Dickens's hand. After Dickens's death, Collins adapted the play as a novella for use in his public reading tour in 1874 (Brannan, 1966). *No Thoroughfare* was also written first as a drama and then rewritten into a novella form.

When the collaborative pieces are divided into smaller units like a chapter or act, eight units (including *The Frozen Deep*) are of single authorship, with the remaining six units being a case of mixed authorship.

Table 1 Bibliographic details for the four collaborations between Dickens and Collins				
No.	Date	Title	Part	Authorship
1	1857, 1866, 1874	<i>The Frozen Deep</i>		Collins & Dickens => Collins
2	1857	<i>The Lazy Tour of Two Idle Apprentices</i>	Chapter I	Dickens & Collins
			Chapter II	Dickens & Collins
			Chapter III	Dickens & Collins
			Chapter IV	Dickens
			Chapter V	Dickens & Collins
3	1857	<i>The Perils of Certain English Prisoners</i>	Chapter I	Dickens
			Chapter II	Collins
			Chapter III	Dickens
4	1867	<i>No Thoroughfare</i>	Overture	Dickens
			Act I	Dickens & Collins
			Act II	Collins
			Act III	Dickens
			Act IV	Dickens & Collins

Fig. 1:

3. Testing the authorship of collaborative chapters

The following experiments draw on a Dickens corpus comprising 22 texts and a Collins corpus with the same number of texts as a basis of reference, with which we compare the style of the collaborative chapters (see Tables 3 and 4 in Appendix for details). The first round of analysis is to run Craig's version of Burrows's *Zeta* test in order to extract Dickens markers as well as Collins markers. The vast majority of authorship attribution studies have relied on the most common words in the corpus/text in question (Burrows, 2002/2005; Eder, 2010; Eder & Rybicki, 2009; Hoover, 2003/2004; Rybicki, 2009; to name but a few). The recent works by Craig & Kinney

(2009) and Hoover (2010; 2011; 2013), on the other hand, have demonstrated Zeta's strong power of differentiation between two sets of text samples. Other keywords extraction techniques include the use of Log-likelihood ratio (Dunning, 1993) criticised by Tabata (2012) for being prone to burstiness (and for its tendency to produce too many false-positives); Mann-Whiney U test (Kilgariff, 2001); t-test (Hoover, 2010); bootstrap test (Lijffijt et al., 2012); Random forests (Tabata, 2012), and so forth. The particular strength of Craig's version of Zeta analysis is its simplicity and effectiveness well documented in Hoover's studies mentioned above. A Zeta distinctiveness ratio for the word *i* is calculated in the following formula:

$$\text{Zeta}_i = \frac{DF(x)_i}{N(x)} + \frac{N(y) - DF(y)_i}{N(y)}$$

Fig. 2:

where $N(x)$ = Total number of text-segments in the corpus x ; $DF(x)_i$ = Document (or Segment) frequency for the word i in the corpus x ; $N(y)$ = Total number of text-segments in the corpus y ; $N(y) - DF(y)_i$ = Document (or Segment) frequency negative for the word i in the corpus y .

The R package *stylo* (a suite of tools for stylometric analyses) (Eder, Rybicki, and Kestmont, 2013) includes the function *oppose()* with an option to calculate Zeta scores. In the present case, each text was sliced into 10,000-word segments with rare occurrence threshold set to 2 so as to exclude hapax legomena and filtering threshold set to 0.5 to extract strongly discriminating words (thus only $\text{words}(i)$ with $\text{Zeta}(i) < 0.5$ or $\text{Zeta}(i) > 1.5$ are picked up). The procedure detected 122 authorial markers: 61 Dickens markers and 61 Collins markers as listed in Table 2.

Table 2 122 Author markers in the descending order of distinctiveness (Filter threshold = 0.5)

Dickens markers:
upon, and, so, very, but, much, great, though, a, being, indeed, many, down, such, or, with, several, scarcely, they, up, off, then, often, shaking, length, observed, honour, towards, short, said, were, beside, fire, there, head, rather, always, forth, glancing, air, looking, afterwards, presently, would, coach, would, countenance, returned, eye, where, pretty, within, well, boy, fact, their, gentleman, nor, should, its, legs
Collins markers:
to, first, words, on, left, only, answered, letter, interests, second, discovered, next, asked, met, future, own, answer, wait, moment, end, serious, position, happened, tried, waited, written, back, me, opened, woman, failed, resolution, later, time, produced, enough, experience, chance, followed, question, return, in, leave, placed, suddenly, questions, she, waiting, room, privately, plainly, my, useless, spoke, writing, discovery, write, her, offered, motive, view

Fig. 3:

When the marker words were fed into a cluster analysis, the resulting dendrogram (Fig. 1) clearly differentiated the two authors in the distinct clusters. The marker words are also sensitive enough to show sub-patterns: both in Dickens and Collins clusters, the texts written in their early career flock together (Dickens's texts in the 1830's and Collins's texts published in the 1850's). Dickens's early works branch close to the sketches/travelogues as opposed to fictions.

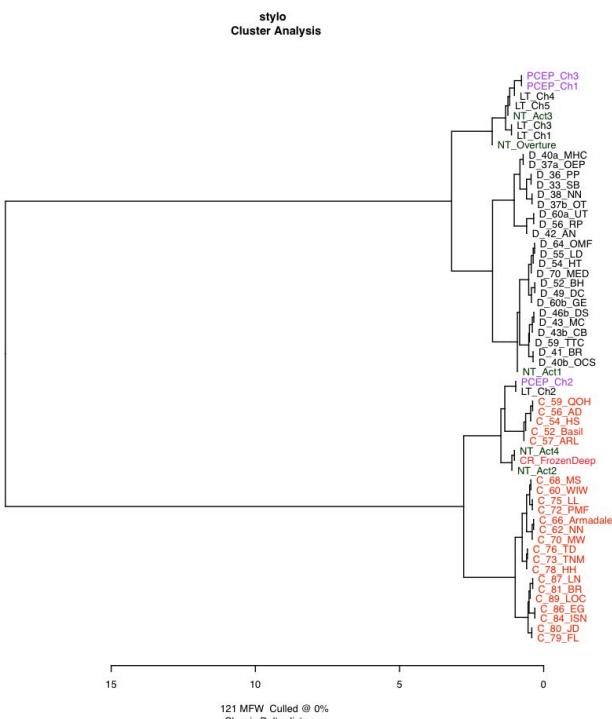


Fig. 4: Cluster analysis: Dickens versus Collins (Distance: Burrows's Delta)

Fig. 4 is a cluster dendrogram with the collaborative pieces included in the analysis. The prominent feature of this result is that:

1. The single-handed chapters/acts fit in well with the author's cluster (Chapters 1 and 3 of *The Perils of Certain English Prisoners* perch in the topmost sub-cluster together with the Overture and Act 3 of *No Thoroughfare*, whereas *The Frozen Deep* and Act 2 of *No Thoroughfare* are found as the nearest neighbours to each other, with Chapter 2 of *The Perils* placed in the Collins cluster).
 2. The co-authored chapters/acts form slightly distant sub-clusters in each of the two author's main clusters

4. Letting the Delta roll through to find dynamic shifts in style

Although the result of cluster analysis, being in consonant with the bibliographical details given in Table 1, helps confirm the effectiveness of style markers found through a Zeta test, it inevitably tells us about the limitation of the procedure that captures only a static snapshot of stylistic similarity or difference between texts. Language is never monolithic throughout a text. Language in fiction varies wildly from narratives to dialogue, or vice versa. Language in fiction is indeed quite mobile.

We need, therefore, a technique to capture dynamic style change. If the collaborated chapters can be sliced into consecutive segments of n words with a partial overlapping so that we can roll through to focus upon a certain stretch of text like a moving camera rather than take the entire text/chapter in one snapshot, it will be possible to detect subtle fluctuations of style between one stretch and another as well as to pinpoint where one author takes over from another, etc. Eder, Rybicki, and Kestemont's *Rolling Delta* was developed exactly for this purpose.

Figure 5 shows a result of Rolling Delta run with a window size set to 3,000 words, a step size of 300 words, using 100 most common words as variables. The whole text of *The Perils of Certain English Prisoners* is cut into consecutive 3,000 word-segments, with each segment compared with the centroids of Dickens and Collins, respectively.

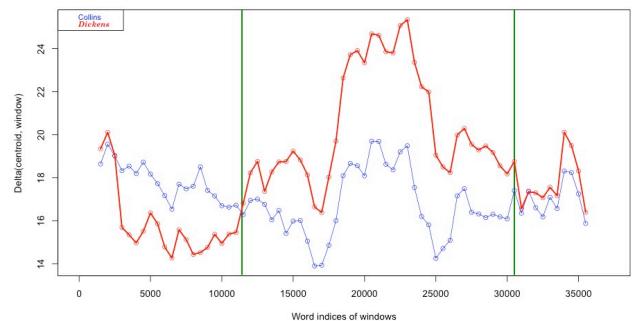


Fig. 5: Rolling Delta: The Perils of Certain English Prisoners and the centroids of Dickens and Collins

What strikes us is that the two major intersections (marked with green vertical lines) roughly correspond to the chapter boundaries in the text, a result that illustrates how the technique is capable of pinpointing possible authorial takeovers.

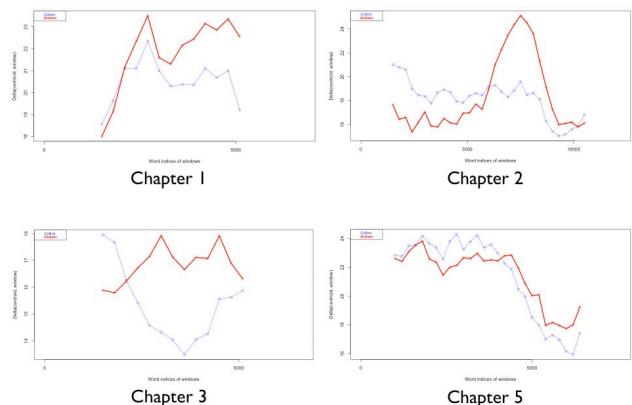


Fig. 6: Rolling Delta: The collaborated chapters of the Lazy Tour (The thicker line indicates Dickens's centroid)

Fig 6 displays Delta polygons of the four collaborated chapters of the *Lazy Tour of Two Idle Apprentices*. Of particular interest is that a remarkably similar pattern holds throughout the four diagrams: it is always Dickens who takes the lead at the outset of a chapter. He runs quarter or halfway (at most) into each chapter before passing over to Collins. The series of Rolling Delta plots seem to reflect an interesting nature of collaboration as well as the unequal partnership (Nayder, 2002) between Dickens and Collins: Dickens always takes initiative, sets a keynote for the whole chapter, which Collins takes over and continues the rest of chapters, a typical relationship between a master and his disciples.

References

- Allingham, P. V.** (2011). *A Comprehensive List of Dickens's Short Fiction, 1833-1868*. The Victorian Web. [Online] (Last accessed 31 October 2013.) <http://www.victorianweb.org/authors/dickens/pva/5.html>

Brannan, R. L. (1966). *Under the Management of Mr. Charles Dickens: His Production of "The Frozen Deep"*. Ithaca, New York: Cornell University Press.

Burrows, J. (2002). *Delta: a Measure of Stylistic Difference and a Guide to Likely Authorship*. LLC, 17. 267-287.

Burrows, J. (2005). *Who wrote Shamela? Verifying the Authorship of a Parodic Text*. LLC, 19/4: 453-475

Burrows, J. (2007). *All the Way Through: Testing for Authorship in Different Frequency Strata*. LLC, 22/1: 27-47.

Craig, H. and A. Kinney (eds.) (2009). *Shakespeare, Computers, and the Mystery of Authorship*. Cambridge: Cambridge University Press.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence, Computational Linguistics, 19/1: 61-74.

Eder, M. (2010). *Does Size Matter? Authorship Attribution, Small Samples, Big Problem*, Digital Humanities 2010 Conference Abstracts, King's College London, 132–5.

Eder, M. and J. Rybicki (2009). *PCA, Delta, JGAAP and Polish Poetry of the 16th and the 17th Centuries: Who Wrote the Dirty Stuff?*, Digital Humanities 2009 Conference Abstracts, University of Maryland, College Park, 242–4. Eder, M., J. Rybicki, and M. Kestemont (2013) Computational Stylistics [Online] (Last accessed 31 October 2013.) <https://sites.google.com/site/computationalstylistics/home>

Eder, M., M. Kestemont and J. Rybicki (2013). *Stylometry with R: a suite of tools*. Digital Humanities 2013 Conference Abstracts. University of Nebraska-Lincoln, NE, 487–9.

Hoover, D. L. (2003). *Multivariate Analysis and the Study of Style Variation*, Literary and Linguistic Computing, 18/4: 341–360.

Hoover, D. L. (2004). *Testing Burrows's Delta, Literary and Linguistic Computing*, 19/4: 453–475. Hoover, D. (2010). Teasing out authorship and style with t-tests and Zeta. Digital Humanities 2010 Conference Abstracts, King's College London, 168–70.

Hoover, D. (2011). *The Tutor's Story: a case study of mixed authorship*. Digital Humanities 2012 Conference Abstracts, Stanford University, Stanford, CA, 149–51.

Hoover, D. (2012). *The rarer they are, the more they are, the less they matter*. Digital Humanities 2012 Conference Abstracts, Hamburg University, Hamburg, 218–21.

Hoover, D. (2013). *Almost All the Way Through — All at Once*. Digital Humanities 2013 Conference Abstracts, University of Nebraska-Lincoln, NE, 223–6

Kilgariff, A. (2001). *Comparing Corpora*. International Journal of Corpus Linguistics 6 (1): 1–37.

Lane, M. (1956). *Introduction. In The Oxford Illustrated Dickens, Christmas Stories*. Oxford: OUP

Lijffijt, J., T. Säily, and T. Nevalainen (2012) *Chi-square test considered harmful: Better methods fo testing the significance of word frequencies*, A paper presented at ICAME 33, 30 May–3 June 2012, Leuven, Belgium.

Nayder, L. (2002). *Unequal Partners: Charles Dickens, Wilkie Collins, and Victorian Authorship*. Ithaca/London: Cornell UP.

Rayson, P. and R. Garside (2000). *Comparing Corpora Using Frequency Profiling, Proceedings of the Workshop on Comparing Corpora*, Held in Conjunction with the 38th Annual Meeting of the Association for Computational Linguistics (ACL 2000), 1–8 October 2000, Hong Kong. 1–6. Available online at <http://www.comp.lancs.ac.uk/computing/users/paul/phd/phd2003.pdf>.

Rybicki, J. (2009). *Translation and Delta Revisited: When We Read Translations, Is It the Author or the Translator that We Really Read?*, Digital Humanities 2009 Conference Abstracts, University of Maryland, College Park, 245–7.

Rybicki, J. and M. Eder (2011). *Deeper Delta across genres and languages: do we really need the most frequent words?*, Literary and Linguistic Computing, 26/3: 315–321.

Stone, H. (ed.) (1968). *Charles Dickens' Uncollected Writings from "Household Words" 1850–1859*. 2 vols. Bloomington: Indiana UP.

Tabata, T. *Approaching Dickens's Style through Random Forests*, Digital Humanities 2012 Conference Abstracts, University of Hamburg, Germany, 388–91.

Thomas, D. A. (1982). *Dickens and the Short Story*. Philadelphia: University of Pennsylvania Press.

paola.andriani@gmail.com

Dipartimento di Filologia, Letteratura e Linguistica, Università di Pisa

Bartalesi, Valentina

valentina.bartalesi@isti.cnr.it
ISTI-CNR

Locuratolo, Elvira

elvira.locuratolo@isti.cnr.it
ISTI-CNR

Meghini, Carlo

carlo.meghini@isti.cnr.it
ISTI-CNR

Versenti, Loredana

loredana.versenti@isti.cnr.it
ISTI-CNR

1. Introduction

In the field of digital humanities, scholars are increasingly producing digital editions of texts and manuscripts. The representation of knowledge included in literary texts is a complex issue, requiring rich vocabularies, also called ontologies, for representing the many different aspects that are investigated by scholars. In literature, there are many ontologies that focus on different aspects of textual information but one single ontology representing all these aspects does not exist.

The “Towards a Digital Dante Encyclopedia” project is a three years Italian National Research Project, started in 2012, that aims at building a prototypical digital library endowed with services supporting scholars in creating, evolving and consulting a digital encyclopedia of Dante Alighieri and of his works. The digital library is based on a semantic representation of Dante's works and of the knowledge embedded in them in RDF language¹, a language recommended by the Web Consortium for the representation of knowledge. In RDF, every piece of knowledge is represented as a triple (subject predicate object), and a set of triples form an RDF graph, generally called semantic network, in order to highlight the formal linguistic nature of the representation. The services being developed address several tasks carried out by the scholars building the encyclopedia, starting with the visualization of references to primary sources (i.e., other authors' works which Dante referred to his own works), their types and their distribution both in time and in the works of Dante. The overall goal is to shed light into the cultural context in which Dante wrote his works and into the development of Dante's reference library over time.

This part of the project is divided in several phases. The first phase regards the creation of an ontology for the knowledge embedded in scholarly commentaries to Convivio², the philosophical treatise which we choose as initial case study. In the second phase, the ontological model is generalized to represent the knowledge embedded in the scholarly commentaries to other Dante's works. In the third phase, Dante's works along with their attached commentaries are inserted into the digital library, as part of the semantic network being built. In the fourth phase, the primary sources referenced by Dante in his works, as reported by the commentaries, are inserted into the digital library, following the same semantic approach. In the last phase, services are developed, as web applications that allow scholars to browse the semantic network of Dante's work, of primary sources, or of references linking the former to the latter. The references will be visualized in an intuitive way through tables and charts, highlighting their distribution in Dante's work and over time.

We present the structure of the semantic network, as it currently stands and indicate how it will be further developed. Furthermore, we highlight the benefits brought by the visualization service of primary sources to scholars.

2. Ontology for the representation of convivio

In order to detect the primary sources used by Dante to write his Convivio, we relied on the most recent and updated commentary to the text, that of Gianfranco Fioravanti

Towards a Semantic Network of Dante's Works and Their Contextual Knowledge

Tavoni, Mirko

mirko.tavoni@gmail.com

Dipartimento di Filologia, Letteratura e Linguistica, Università di Pisa

Andriani, Paola

(Mondadori, in the press), and created an ontology for representing the relevant knowledge carried by this commentary. In particular, our ontology represents:

- the passage of Dante's text (e.g., "Si come dice lo Filosofo nel principio della Prima Filosofia") to which a quotation from a source refers;
- the correspondent book, chapter and paragraph of Dante's text
- the author of the work referenced in the commentary (e.g., Aristotle);
- the title of the work referenced in the commentary (e.g., Metaphysics);
- the thematic area of the work referenced in the commentary (e.g., Aristotelianism).

In order to create an ontology for the semantic representation of the above information, we investigated several existing ontologies (e.g. CIDOC-CRM³, FRBR⁴, FaBiO⁵, SKOS⁶), and we chose the classes and properties that we considered the most appropriate to represent the above information. Furthermore, we added our own classes and properties for the representation of the categories of knowledge that were not addressed by the existing ontologies. Then, we transformed the initial commentary into an RDF graph structured according to the ontology⁷.

On the basis of our ontology, we are approaching the remaining phases of the "Towards a Digital Dante Encyclopedia" project. To such aim, we are using the ontology developed so far in order to represent other works of Dante (e.g. *De Vulgari Eloquentia*, *Monarchia*) as well as the knowledge carried by commentaries to them. At the same time, we are collecting the primary sources of Dante's work in a digital format, for insertion into the semantic network underlying the digital library. Our diachronic analysis, in fact, aims at representing the evolution of Dante's knowledge about primary sources.

3. The model population

In order to enrich our RDF graph, as we have done for Convivio, we are collecting information for other Dante's works. In particular we are focusing on (i) the text of the work along with the attached commentaries; (ii) the primary sources referenced in the notes. We are currently storing the RDF triples generated according to our ontology both for the notes and the primary sources. We are relying on the Virtuoso [8] technology for storing and accessing large RDF graphs.

It is important to note that the works of Dante, as well as most of their primary sources, exist in some digital format. However, to the best of our knowledge, there is no semantic representation that integrates this information into a unique body of knowledge, expressed through a formal ontology. We do not expect the knowledge base that we build to give a coherent view of Dante's works. The knowledge in it may, and in general will be incoherent and incomplete, and our ontology is flexible enough to allow both.

The creation of the semantic network is a very time-consuming and knowledge-intensive process. It requires researching the most appropriate ontologies for representing all aspects, and in several cases it requires developing a new ontology to fill the gaps of existing ones. Once the ontology is created, the works of Dante, the primary sources, and the knowledge embedded in them will have to be expressed in this ontology, and this is also a technically demanding task. But the benefits are enormous. The digital representation of the knowledge can support scholars in several conceptually simple but time-consuming tasks, allowing them to focus on the more intellectual aspects of their work. The semantic network will be usable for a wide variety of purposes, which go well beyond the specific services built by our project. It will constitute a backbone that can be enriched with other knowledge about Dante and the historical events, people, artistic movements, etc. that have come across Dante and as such contribute to form the context in which Dante's life and art took place. In this sense, creating the semantic network is the most important achievement of our project. Our project will build only one part

of this network, but will also lay the bases for the extensions and enrichments that will complete what we have started.

4. Why this unified archive will be important to study the culture of Dante

The importance of the archive and tools described above in order to study how the culture of Dante developed in time is obvious. The fact of gathering the current information on the primary sources used by Dante in his works, and the fact of having this information available in digital format, will improve and make more efficient the research of primary sources by the scholars. Having all the information dispersed on paper books, in fact, makes impossible a systematic overview of the culture of Dante and a well-ordered perception of how it was gradually set up in time. On the contrary, the automatic visualization of data about primary resources, according to different parameters (in chronological order, or by type of source, or by author, by work, etc.), will allow to explore the dynamics of the multi-faceted culture of Dante in relation to the diverse and often conflicting stages of his biography and to study the evolution in time of Dante's cultural background.

References

1. **Manola F., Miller, E.** (2004). *RDF Primer*. W3C Recommendation 10 February 2004. Available at: www.w3.org/TR/rdf-primer/
2. **Alighieri, Dante.** (1987). *Convivio*, a c. di P. Cudini, Milano, Garzanti
3. **Doerr, M.** (2003). *The CIDOC CRM - an ontological approach to semantic interoperability of metadata*. AI Magazine, vol. 24(3), 75-92
4. **Tillett, B.** (2004). *What is FRBR? A conceptual model for the bibliographic universe*. Library of Congress Cataloging Distribution Service, vol. 25(5), 1-8.
5. **Peroni, S., and Shotton, D.** (2012). *FaBiO and CiTO: Ontologies for describing bibliographic resources and citations*. J. Web Semantics: Science, Services and Agents on the World Wide Web.
6. **Alistair, M., Matthews, B., Wilson M., and Brickley, D.** (2005). *SKOS Core: Simple knowledge organisation for the Web*. In Proceedings of Dublin Core Conference.
7. **Bartalesi, V., Meghini, C., Locuratolo, E., Versenti, L.** (2013). *A preliminary study on the semantic representation of the notes to Dante Alighieri's Convivio*. In Proceedings of the 1st International Workshop on Collaborative Annotations in Shared Environment: metadata, vocabularies and techniques in the Digital Humanities (DH-CASE '13). ACM, New York, NY, USA.

A "Deeply Annotated" Bibliography of Local Social Histories of Early Modern Europe

Theibault, John Christopher

John.Theibault@stockton.edu
Richard Stockton College of New Jersey, United States of America

Issue:

Beginning in the 1950s, social historical investigations of single villages, towns, cities, and regions of early modern Europe emerged as a significant genre of historical writing. Authors such as Pierre Goubert and Emmanuel Le Roy Ladurie demonstrated that the archival information density for early modern European towns and villages was just right for undertaking a kind of "total history" of the locality. These studies made claims to address important historical issues beyond the

locality under investigation (often having to do with the transition to modernity) that were impossible to cover at the level of the nation or Europe as a whole. Though often identified by fairly generic labels such as “social histories” or “micro-histories,” they distinguished themselves from traditional local histories by explicitly promoting their exemplary character. Sources were generally more abundant than in medieval Europe, as parish registers, tax lists, and court protocols became more abundant; but the information was also pre-statistical, unprocessed by its creators. The timing for working with denser local sources was apt because new computer technology in the 1960s allowed historians to build databases that eased analysis of the pre-statistical material. Historians began to use specialized techniques from the social sciences such as Gini-coefficients of inequality and family reconstitution to work with the historical data in order to work with the data from this pre-statistical age.

As a result, early modern European history emerged as one of the historiographically richest historical fields by the 1970s. By the 2000s, however, the initial promise of “total history” from quantitative social science history was in retreat. Some of that retreat may be attributed to conceptual overreach in the first wave of local studies. Some works seemed to make excessively sweeping claims on the basis of limited cases. But perhaps equally consequential is that the sheer numbers of local studies made it hard to gain an overview of what had and had not been investigated for different parts of Europe.

Current Implementation:

The Early Modern European Social History Geospatial Bibliography Project (EMESHGB) is designed to help transcend these limitations and potentially spur new research in local histories by demonstrating the achievements of earlier research and identifying new paths to explore. It operates on several levels. At its simplest, it is a database of monographs about the localities of early modern Europe written since the 1950s. That database is enhanced by being accessible immediately through a map-based interface with a temporal slider. It is the first systematic effort to show the relationships between works concerning different parts of Europe, so that one can quickly establish what regions have been studied in depth and which are relatively under-researched. But the database will contain additional layers of information that can also be accessed geospatially and temporally from the interface. It is these additional layers that I call “deep annotation.” The deep annotation in the database is a more complex process. Some of it can be carried out without access to the books, directly from MARC records, but other parts of it require access to the works themselves. So far, the development of the ontology of the database and the required vocabulary for search is being done as an iterative process by looking at the individual works. In this early phase, it is limited to monographs written in English about all parts of Europe, along with a small subset of monographs written in French about France and in German about Germany, with the intention of expanding the geographical and linguistic scope once the robustness of the database ontology is clear.

The purpose of the deep annotations will be to group information contained in each work in ways that will facilitate comparisons. The database will contain a full bibliographic citation. It will also extract information easily from that bibliographic citation that users would especially wish to search by: 1) date of publication, 2) range of historical dates covered in the book, 3) name of the locality covered, current country of locality, 4) type of locality covered (e.g. village, town, seigneurie, neighborhood of a city, historic province). From that information, the user could locate, for example, all village studies that cover the years from 1550-1600. The value-added annotations in the database will make even more complex comparisons possible. The key ones will be 1) principal sources used for analysis (e.g. local court protocols, parish registers, cadastral records, tax lists, personal correspondence), 2) social groups analyzed (e.g. peasants, nobles, artisans, burghers, women, children, outcasts), 3) social history concepts invoked (e.g. historical demography, proto-industrialization, inheritance practices, rebellion and resistance), and 4) social science methods employed (e.g. family reconstitution, transition

matrices, Gini coefficients of inequality, total factor productivity). These categories are populated not by locating every possible mention of a technique or source, but by identifying those that figure most prominently in the overall arguments of each work.

There is yet another way in which the geo-spatial and temporal information in the database will allow for a more complex visualization of historiography. Some of the works cited in each individual work will be other works in the database. There will, of course, also be many works cited that are not in the database. The ability to visualize citation networks will help establish which works have been most influential within and across national research traditions.

The linkage of a database with a geo-spatial/temporal interface is no longer that rare in digital humanities. There is also a precedent for the database in question being on its most basic level a bibliography. For example, the Perseus Project located classical works on a spatio-temporal interface. Nevertheless, the EMESHGB is innovative in allowing users to isolate thematic elements in the bibliography for comparison on the map interface. It will prove an important new tool for the next generation of research into Europe’s transformation to modernity. Researchers interested in undertaking their own local studies and wishing to maximize their impact will be able to quickly assess what kinds of questions have been addressed using what kinds of materials for what parts of Europe. They can examine under-researched regions, identify key questions that might usefully be applied to a different geo-spatial region, and determine how their study might be innovative within the genre. Researchers interested in the comparative development of Europe will be able to quickly identify comparable studies for cross-cultural comparisons.

Future Directions:

The first phase of building out the database of the bibliography relies on the subject matter expertise of a single scholar. The initial set of monographs is large enough to provide a useful test of the concept of the database and interface with a complete collection of information, without being so large that it cannot be processed. One reason for beginning the project “by hand” is because it is useful to check the structure of the database against its output with works that are familiar. Also, the fact that the works included in the database are almost all still under copyright has limited the opportunities for larger scale corpus analysis. However, there are opportunities for automating the assignment of information to the different fields in the database by means of topic modeling, the results of which can then be compared with those produced by the scholar. At the same time, we will be establishing a process to transition the project from a “one-off” single development team product to a permanently extensible project. That process will almost certainly involve some kind of tool for allowing people to contribute to the database directly. The project can be extended in several segments after the initial concept has been demonstrated. With each extension, the original database will gain additional utility. After completing the English-language monographs, the most obvious extension is to cover the foreign-language monographs from the same time period. After completing that phase, which would have to be done in some collaborative format with scholars in Europe, we can consider extending the project on one of several dimensions – chronologically (e.g. adding works addressing the medieval period), geographically (e.g. adding works on Latin America and North America in the early modern era), or genre of historical writing (e.g. adding journal articles).

**What remains to be done –
Exposing invisible collections in the
other 6500 languages and why it is
a DH enterprise.**

Thieberger, Nick
thien@unimelb.edu.au
 University of Melbourne

Introduction

In a recent overview of issues in the digital humanities, Manfred Thaller notes the importance of: "(1) access to the information needed to tackle a research question, (2) the analysis of that information by tools reflecting the methodological requirements of the specific discipline and research problem and (3) the publication of the new information gained by the analytical process." (Thaller 2012:11)¹

For most of the world's 7000 languages there are few records available via the internet. Efforts to increase the documentation of these small languages have led to the development of tools and repositories over the past decade. I suggest that Thaller's desiderata are reflected in the language documentation activities of the creation of archives, metadata systems and the ability to locate, store, retrieve and re-use language records. The network of language archives represented by the Open Language Archives Community (OLAC) has adopted a common metadata system that each archive serves for OLAC's aggregation, allowing more specific searches than can be provided by google, for example. However, not all digital language archives currently provide metadata to OLAC, rendering their collections invisible to the aggregated search. While their webpages may be accessible to web-searches, they do not allow the targeted search by language that is the focus of OLAC's aggregator. Other repositories (including many institutional repositories—national libraries and archives, mission archives and so on) have language content that is not noted in the collection's catalog, and the catalog may not be available for web-harvesting. Finally, there are collections still held by their creators and not in a repository at all.

This paper discusses two approaches to making collections of primary language material locatable and accessible. While the methods are generalisable to any discipline, this paper describes an index of records of language material for collections that have no such metadata and for which no other mechanism is foreseeable. The first approach builds a traceable index of a researcher's discoveries in existing repositories, for example, a state library or archive, using established aggregation services. The second is a survey that aims to locate and digitise smaller collections that are currently outside established institutions, typically still in the care of the researcher.

The language index

The language index provides metadata in an Open Archives Initiative-compliant form, allowing records² to be found in generic language searches³. Not all repositories can provide metadata using ISO-639-3 language codes, so it is useful to provide a mechanism whereby researchers can build this resource as they discover new material. In general, repositories are just unaware of standards rather than being reluctant to share data, hence the need for them to either change their metadata system (which is unlikely) or for an index of the kind described here, that points to their collections.

The paper will demonstrate the index as it is currently implemented in the catalog of the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC), and will discuss the issue of persistence of the links provided and future possible alternatives to exposing collections that are otherwise invisible to aggregation.

The survey

In an effort to locate what I have called invisible collections I launched a survey⁴ in 2012 asking respondents to identify language collections that needed to be either (or both) digitised or described using OLAC's service. I tried to keep the questions

as simple and easy to answer as possible. In this paper I will report on the findings of this survey, in particular noting that they point to the need for training; for simple metadata entry tools; for standards-compliant metadata repositories; and for recognition of collections of primary material as a form of scholarly output.

The survey questions were as follows:

1. Do you know of recordings of small or endangered languages that are not yet digitised? These could be in personal collections or in established repositories that do not plan to digitise their collections. If so, please provide as much detail as you can about the number and type of recordings (reel to reel, cassette, DAT etc), the content, and the state of their current storage. Can you provide information about who to contact about these collections?

2. Do you know of collections whose catalogs are not available through federated searches (that is, they are only available if you visit their website and not anywhere else on the web) and for which we could provide a reference to make it easier to find them?

3. Do you know of repositories of manuscripts that have received little attention from linguists but which are likely, in your opinion, to have linguistic records in them? These may include, for example, missionary archives or State administrative archives.

4. Please include your name and contact email so we can follow up with you if necessary (email addresses will not be added to any lists). (Please indicate if you allow us to publish an anonymised version of your response).

The survey form was publicised among linguistic networks. It is now nominated as a future activity of the international network of language archives, DELAMAN⁵ which should ensure wider coverage. As a first step, it has revealed an interesting variety of collections, each with characteristics that are significant for the effort of making such collections available. At a time when funding for digitisation is difficult to obtain, it is important to recognise that unique cultural heritage recordings such as these are at risk of being lost. A summary of some responses and an observation about the broader significance of each is given below.

(1) 22 tapes of a Sudanese language held in Washington DC by a retired linguist – how to get them digitised and where to store them then? 22 tapes are sort of manageable. There are also a large number of notes that need to be scanned. For a retired researcher it may not be easy to access the equipment needed to do this work.

(2) Several hundred cassettes in a Solomon Islands language, particularly valuable as they are recorded by a speaker, so capturing lots of natural speech. Digitising such a collection is a serious undertaking needing significant funds.

(3) The tapes are in Stockholm, stored in a box but the recorder is based in Chicago and is still an active academic. A basic problem of access of the collector to their own material.

(4) Colorado, USA, a dozen reel-to-reel and two dozen cassette tapes with a senior linguist concerned to make the collection safe and not being sure what to do.

(5) Tapes were deposited with a national Cultural Centre in a small Pacific country that may or may not have the resources to look after them. It does not publish its catalog (if it actually has one) so it is not clear if these tapes need to be digitised or not, or what conditions may be placed on access to them.

(6) A recent MA in Linguistics at one of the PARADISEC consortium universities, tapes stored in boxes. Paper transcripts may have been thrown out. Shows lack of communication even within our own departments.

(7) A collection of [language] tapes stored in a Harvard University repository which may not prioritise digitising it (but could if funding were made available).

(8) Researchers who have small collections, less than twenty tapes, and digitise them themselves by connecting a tape player to a digital recorder. Problem of methods used in digitisation, may damage the tape and not result in the best digital file.

One reason that these collections are not digitised is clearly the lack of importance placed by academia on the re-use of primary research materials. If it were an acknowledged research output to create archived and accessible collections of

primary data, counting towards promotion and tenure, then it is more likely that cases like those listed above would no longer occur. It is clear that much remains to be done to extend the reach of digital language archives, assisting in locating legacy collections, describing and digitising them, connecting with source communities/individuals, creating a means for online annotation (crowdsourcing) and of valuing the collections (both monetarily or academically). I conclude by discussing an online service for providing small metadata snippets pointing to these otherwise invisible collections. This paper presents these efforts based around the digital archive Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC).

References

- Thaller, Manfred** (Ed.). (2012). *Controversies around the Digital Humanities*. Historical Social Research Vol. 37 (2012), No. 3.
www.language-archives.org/item/oai:paradisec.org.au:JL1-link
www.language-archives.org/item/oai:paradisec.org.au:JL1-link
www.paradisec.org.au/PDSCSurvey.html
www.delaman.org

Photogrammar: Organizing Visual Culture through Geography, Text Mining, and Statistical Analysis

Tilton, Lauren
Yale University, United States of America

Leonard, Peter
Yale University, United States of America

Arnold, Taylor
Independent Scholar

The Farm Security Administration – Office of War Information photographic dataset is a collection of over 170,000 monochrome and colour photographs, commissioned between 1935 and 1945 by the government of the United States of America. Offering a unique snapshot of the nation during the period, it serves as an important visual record for scholars and the public at large. The FSAOWI photographic archive has been digitized by United States Library of Congress, and because the photographs were taken on behalf of the United States Government, access to and use of the collection is essentially free and open.

Under the direction of Professor Laura Wexler, (American Studies and Women, Gender & Sexuality Studies, Yale University), the Photogrammar project takes the conditions of this archive (predigitized and publicly accessible) as the starting point for a digital, scholarly and open source platform that builds upon, and significantly extends, the Library of Congress' online collection. The subject of a successful Digital Humanities Start-Up Grant from the United States National Endowment for the Humanities, the project seeks to answer research questions that emerged from scholars at Yale University. Our paper will focus on three Digital Humanities techniques that we have used to analyze the corpus:

Geospatial: Computational derivation of latitude and longitude;

Text Mining: Vectorspace analysis to expose thematic similarity;

Statistical Analysis: Contextual inference to “rediscover” missing metadata.

Each of these techniques is associated with significant information gain, relative to the previous state of scholarship on the corpus:

Geographic: The collection is often characterized as being about the dust bowl and rural poverty in the American south during the Great Depression. In fact, by mapping the

photographs and analyzing photograph density at the county level by year, the popular characterization of the FSAOWI does not hold. Rather, the scope of sight was much broader including a large focus on the United States Northeast and Midwest, as well as photographs beyond the continental United States such as the Virgin Islands and Europe.

Text Mining: The 1940s ontology of the collection only allowed a photograph to be classified in one category at once. By looking at latent patterns in the freeform textual descriptions, we are able to surface photographs that participate in multiple and overlapping clusters. In this way, we can discover thematic similarity between the work of several different photographers, active at different times and in different places in the country (and around the world). This approach reflects a more general turn towards ‘latent’ patterns in unstructured data within the archive.

Statistical Analysis: The relatively large scale of the collection (available as both digitized negatives and physical prints) as well as constantly changing organizational systems through the years has unfortunately left a majority of the negatives with minimal documentation. Utilizing latent metadata attached to the photographs, we are able to take individual photographs and to put them back into strips of four and five. In turn, this allowed for us to insert new metadata into the photographs. For example, if a frame with an unknown photographer and location is between photos by John Vachon in Chicago, then we know the unknown frame is by this photographer in this location. We will discuss the statistical methods applied using R.

These three techniques open up new questions about this collection and historic period, and challenge previous scholarship. We believe the Photogrammar project can serve as one example of the general question of how to engage with largescale digital archives of visual culture. This question is of particular importance for scholars who seek to bring Digital Humanities techniques to “Big Data” collections, whether those curated by libraries, museums, or scholars themselves. We anticipate both similarities — and important differences — with European archives of the same period, including the UK Mass Observation Archive (1937-1960s), and forthcoming collections hosted by Europeana Online.

In addition, we will discuss how this project offers a new, userfriendly way to access a visual archive of this size by sitting at the intersection of *public* and *digital* humanities. We will discuss the ways in which we intend to open up the collection to contributions from the public at large, with lessons learned from previous attempts to crowdsource metadata for this collection (Flickr Commons / New York Public Library 20084, Flickr Commons Library of Congress 20095). We will show a prototype of a publicallyaccessible Geographic Referencer, to allow end users to more accurately and appropriately locate photographs on both current and historic maps. And we will discuss some of the challenges in incorporating crowdsourced metadata corrections into a historic archive, while preserving the integrity and historic character of a large visual collection.

References

- <http://www.loc.gov/pictures/collection/fsa/>
- http://www.loc.gov/rr/print/res/071_fsab.html
- <http://americanstudies.yale.edu/faculty/laurawexler>
- <http://www.flickr.com/photos/nypl/sets/72157610969038056/>
- http://www.flickr.com/photos/library_of_congress/sets/72157618541455384/

A novel approach for a reusable federation of research data within the arts and humanities

Tobias, Grädl
tobias.grädl@uni-bamberg.de

University of Bamberg, Germany

Henrich, Andreas

andreas.henrich@uni-bamberg.de
University of Bamberg, Germany

1 Introduction

In distributed systems literature the orthogonal but interdependent characteristics of *autonomy*, *distribution* and *heterogeneity* are used to classify distributed systems [1,2]. From a holistic perspective on the arts and humanities, collections have evolved over decades or centuries from highly autonomous disciplines and institutions and are widely spread, which resulted in heterogeneous perspectives and data models [3]. Despite its negative notion as *data integration problem*, the term *heterogeneity* also symbolizes the diversity of research methodologies within the disciplines of the arts and humanities. Resolving heterogeneity hence implies an abstraction from the specifics that are valuable for focused disciplinary and interdisciplinary research projects.

Our approach presents a novel concept for data federation in the arts and humanities, which focuses the needs of research projects as well as interdisciplinary and broad use-cases. We especially address the reusability of explicated knowledge on correlations between schemata and digital collections and show where domain experts are required to bridge semantic gaps.

2 Background

Approaches to data integration often follow the theoretical foundation expressed in [4] by employing the concept of a global view. While being highly distinctive in terms of their underlying concepts, established examples such as ISIDORE [5], OAster [6] and Europeana [7] share the goal of facilitating access to a wide range of research data through integrated schemata or ontologies. Aside from broad services, an integration need that focuses on a specific topic and related research questions is addressed by the Steinheim Institute, which provides a search in the context of german-jewish history and judaism [8]

Despite usability concerns in having to identify relevant services and accordingly collections, the reappearing need to overcome the same aspects of heterogeneity in reaction to new use-cases is one of the problems we address.

2.1 Use cases

Our federation concept is primarily focused on the realization of the *DARIAH-DE Generic Search* (<http://dev3.dariah.eu/search>), which includes support for queries over large sets of unrelated collections (broad search) and tightly correlated data (deep search)—with different information needs in mind:

Broad search: Due to the quantity and distribution, the relevance of digital collections for particular research questions is not easily assessable. A broad view assists scholars in finding and evaluating possibly relevant data. Figure 1 shows an exemplary, collection-level aggregation of results based on term statistics in the prototype of our search, which will be continuously extended by other relevant visualization techniques (e.g. with respect to spacial and temporal aspects).

- Deep search: If the granularity of local data models can be used to formulate more specific queries targeting structure and content (e.g. in search facets), the deep search utilizes mappings specified in the *DARIAH-DE Schema Registry*. Broad search continuously fades to deep search with an increasing count and richness of mappings and hence typically smaller sets of selected collections.

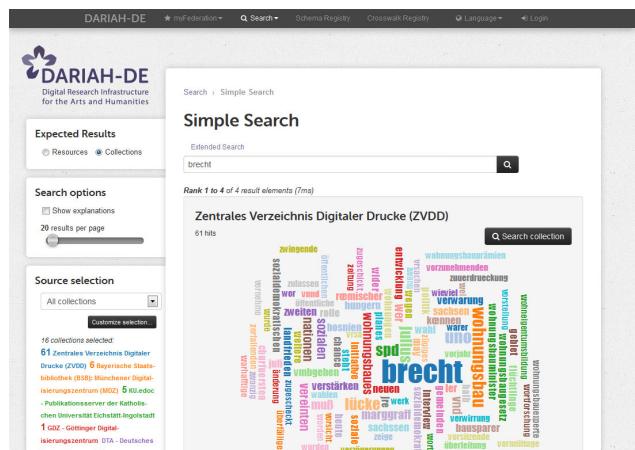


Fig. 1: Result aggregation in the generic search

Despite the focus on the generic search with its virtual integration at query-time, the proposed concept also addresses requirements of a materialized integration of data:

- Data migration and consolidation: Traditional applications of data integration often do not require a dynamic adaption to selected collections, but determine a set of relevant data sources and an appropriate integration schema or ontology [4]. Examples include data migration induced by the introduction of new information systems (e.g. replacement of outdated archive information software) or the consolidation of selected data sources under a merged schema for the purpose of interdisciplinary analysis and visualization e.g. in the *DARIAH-DE GeoBrowser* [9].

2.2 Problem Definition

The common objective of data integration approaches is to resolve heterogeneity on various levels: *Syntactical* aspects such as the existence of different access and encoding methods can be solved by technical means, whereas *structural and semantic* heterogeneity depend on the application of background knowledge [1]. Despite continuing efforts in the fields of schema and ontology matching, the manual intervention of domain experts—especially for large or complex schemata and ontologies often found in the arts and humanities—has shown to be essential to generate high-quality results [10].

[...]. The correlation of the used schemata and ontologies is an inherently complex manual task in our context, which depends on the *fragmented and distributed knowledge* of individual disciplines, collections and scholars. Requiring a common understanding, research projects concentrate knowledge about schemata and semantics used in relevant collections and specify meanings and correlations. In order to integrate the described data and establish technical interoperability, an application of digital methods and tools is required.

3 Concept

Abstracting from aspects of technical and syntactical heterogeneity concerned with accessing, preprocessing and integrating data in a generic fashion, we aim to enable researchers to focus on those aspects of integration, that depend on their knowledge and expertise: the description and correlation of schemata and ontologies. Despite the immediate benefit for individual integration tasks, the centralized formalization and explication of semantics results in the significant advantage of *knowledge reusability*.

3.1 Semantic cluster

The logical architecture of our idea is represented by a directed, weighed graph, where the schemata and ontologies are described by vertices, and mappings between them are

symbolized by edges. Whereas correlations between structural elements symbolize a relation of the described concepts (e.g. persons, locations) and could be considered undirected, more specific rules that are required for data transformation can be composed of non-reversible functions (e.g. the concatenation of fields). For that reason, parallel edges are required for the description of both mapping directions. Differences of schemata in terms of their complexity and expressiveness reduce the achievable level of accumulated completeness, which is represented by the value of *cohesion*.

Figure 2 indicates how the cohesion between schemata can be utilized to suggest semantic clusters: C1, C2 and C3 could be the result of research projects, which needed a high level of mapping completeness between relevant schemata. By interrelating clusters or generically used schemata (S10), the expressed semantics can be reused in other contexts.

3.2 Use-case orientation

Our example indicates the difference to the commonly found integration pattern of a global ontology or schema. Despite its theoretical foundation, simplicity and proven applicability for broad integration use-cases [5,6,7], we consider the approach to be impracticable for a holistic context of the arts and humanities because a global structure would either have to be an abstraction from collection or discipline specifics or unmanageably complex.

Narrowing this context to individual domains or research projects, standards could be elected as appropriate integrative structures. As exemplified in figure 2, the schemata S3, S5 and S8 form the integration baseline within our clusters due to their cohesion with other schemata. Considering our *deep search and data migration and consolidation* use-cases, these schemata can be utilized to generate a fine-grained view over selected collections accessible within the cluster. In order to support interdisciplinary use-cases, clusters can be combined (symbolized by the strong cohesion between S5 and S8) to resolve semantic gaps.

For broad use-cases we rely on the collaborative and continuous emergence of schemata or ontologies (compare S10) within our federation that are used to connect the clusters on the coarse levels sufficient for broad use-cases.

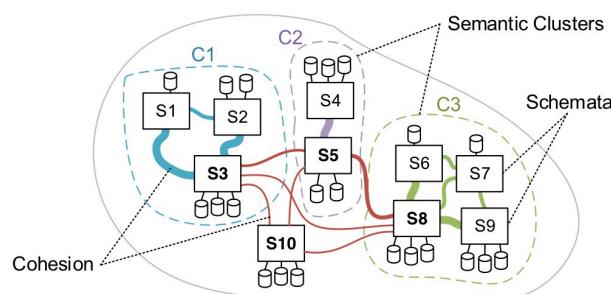


Fig. 2. Semantic clusters of schemata

3.3 Scalability considerations

The simplicity of traditional data integration emerges as new local schemata are added to the system and hence an appropriate mapping target needs to be identified. To ensure extensibility and scalability, our proposed federation concept depends on two strategies:

Cluster globals: The concept of semantic clusters builds on the existence or advancement of standards that are considered as appropriate common perspectives by research communities. Although clusters are not predetermined but expected to evolve, established standards such as the *CIDOC Conceptual Reference Model (CIDOC CRM)* or the *Text Encoding Initiative (TEI) Guidelines* could be identified as initial cluster schemata, which can be mapped in a generic fashion [11]. As new schemata need to be added, the standard which promises to achieve the highest completeness is selected to be mapped.

Model inheritance: Our proposal includes an approach to specify the actual usage of schemata more precisely than it is possible at the level of generic crosswalks. Figure 3 shows the exemplary derivation of the Dublin Core element dc:coverage to resolve an encapsulated substructure. Mappings are inherited to correlate the refined elements or to specify detailed data transformation rules. As derived schemata are related to their parent, generic mappings remain valid and can be utilized if specific rules are missing.

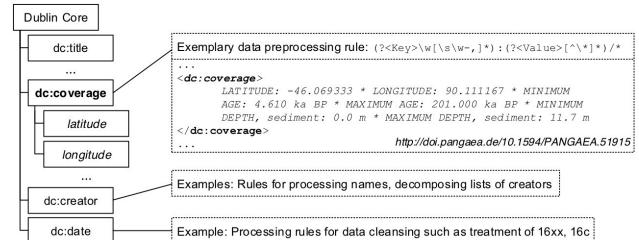


Fig. 3: Exemplary derived version of Dublin Core

4 Conclusion

As we abstract from technical aspects of heterogeneity and reuse the valuable disciplinary knowledge explicated in terms of correlations, processing and transformation rules, the efforts required for integrating research data can be significantly reduced. Another important aspect that is currently being evaluated consists in appropriate techniques for the visualization of our federation concept and system. After all, domain experts need to be able to recognize clusters, important schemata and ontologies as well as their correlations in order to identify semantic gaps and to collaboratively fill them.

References

1. Sheth, A.P., Kashyap, V. (1993): *So Far (Schematically) yet So Near (Semantically)*. In: Proceedings of the IFIP WG 2.6 Database Semantics Conference on Interoperable Database Systems (DS-5), Amsterdam and The Netherlands and The Netherlands, North-Holland Publishing Co 283–31
2. Busse, S., Kutsche, R.D., Leser, U., Weber, H. (1999): *Federated Information Systems: Concepts, Terminology and Architectures*
3. Henrich,A., Gradi,T.:DARIAH(-DE) (2013): *DigitalResearchInfrastructurefortheArtsandHumanities—ConceptsandPerspectives*. International Journal of Humanities and Arts Computing 7(supplement) 47–5
4. Lenzerini, M. (2002): *Data Integration: A Theoretical Perspective*. In Abiteboul, S., ed.: Proceedings of the twenty-first ACM SIGMOD-SIGART-SIGART symposium on Principles of database systems, New York and NY, ACM 23
5. Pouyllau, S. (2011): *ISIDORE : acces to open data of arts & humanities*
6. Hagedorn, K. (2003): *OAIster: a "no dead ends" OAI service provider*. Library Hi Tech 21(2) 170–181
7. Peroni, S., Tomasi, F., Vitali, F. (2013) : *Reflecting on the Europeana Data Model*. IAgosti, M., Esposito, F., Ferilli, S., Ferro, N., eds.: Digital Libraries and Archives. Volume 354 of Communications in Computer and Information Science. Springer Berlin Heidelberg, Berlin and Heidelberg 228–24
8. Lordick, H. (2013): *Vieles finden – die Suchmaschine im Steinheim-Institut*
9. Romanello, M. (2013). *DARIAH Geo-browser: Exploring Data through Time and Space*
10. Rahm, E. (2011): *Towards Large-Scale Schema and Ontology Matching*. In BellahseneZ., Bonifati, A., Rahm, E., eds.: Schema Matching and Mapping. Springer BerlinHeidelberg, Berlin and Heidelberg 3–27
11. Baca, M., Harpring, P., Ward, J., Beecroft, A.: *Metadata Standards Crosswalk*

Problems in Modeling Transactions

Tomasek, Kathryn

s.bauman@neu.edu

Wheaton College, Norton, Massachusetts, United States of America

Bauman, Syd

s.bauman@neu.edu

Northeastern University, United States of America

Introduction

In a paper we presented at the 2012 TEI Members' Meeting[1] and published in the Journal of the TEI[2], we introduced the idea of the "transactionography" as a way to model transactions found in historical financial records (HFRs). In this short paper we briefly review transactionography as a model and discuss a few problems that have arisen as we have been testing it.

Transactionography

We have developed a data structure for recording transactions found in historical financial records and linking that data structure with the segments of running prose or apparently tabular information that attest it or refer to it. While this is certainly a labor-intensive approach, we consider it nonetheless one worth serious consideration.

Our model is similar to the TEI P5 models for contextualization, which offer TEI-compliant models for standoff markup of extra-textual information in files such as prosopographies and gazetteers. Thus in the same way that a prosopography maps how a name of a person that appears in running prose refers to said person, a transactionography maps how the financial records that attest a transaction refer to said transaction. That is, there exists a real-world thing or action that is being referred to by the words (or other marks) in a document we are encoding. The transactionography is a separate file that brings together information that is attested in multiple archival documents to describe a series of transactions.

We noted in our previous paper:

Our model of a transaction is perhaps somewhat more inclusive than some dictionary definitions. We think of a transaction as a coherent set of one or more transfers of something of perceived value from one entity to another. A transfer has three main components, which may be summarized as "what", "from whom", and "to whom". Furthermore, the "what" is likely to be divisible into a certain amount (that is, a quantity and a unit) of a given commodity. It is worth noting that transfers take place at a certain point in time, even if we don't know when that time was. Or, at least, transfers are completed at a certain point in time. A transfer may occupy a significant duration of time from start to finish. For example, a transaction may be conducted by ground postal service.

Some common transaction categories can be identified.

- A standard exchange of money for goods is a purchase, consisting of two transfers: e.g., I transfer \$2 to the convenience store, and a tiny little bottle of apple juice is transferred from the store to me.
- A similar trade that does not involve currency is a barter, for example, trading a large red paper clip for a fish-shaped pen.
- A gift is a single uni-directional voluntary transfer; a single uni-directional involuntary transfer is called theft or embezzlement.
- A set of transfers among more than two entities may be referred to as a multilateral trade.

To model the "what", the TEI <measure> (or <measureGrp>) element seems tailor-made for the purpose. To model the "when", the attributes from the TEI att.datable class seem to be quite reasonable: they can handle specific dates, times, date ranges, and particular dates that took place sometime within a range.

For representing "from whom" and "to whom" we have created new attributes, fra= and til= (Norwegian for "from" and "to") as the attributes from= and to= already occur in TEI, and creditor= and debtor= have different meanings in different contexts.

Problems

Data capture interface

Complex transactions

In our model, a _transaction_ is a series of one or more _transfer_s from one entity to another. (An entity here is a person, an organization, or an account.) This works well for simple purchases, barters, and trades (each of which comprises 2 transfers), as well as gifts and theft (each of which comprises 1 transfer). It also works well for simple multilateral trades. E.g., if Mr. Baxter gives 86.50 USD to Mr. Sheldrake, who gives flowers to Ms. Kubelik, who gives basketball tickets to Mr. Baxter, the transaction comprises three transfers.

Our model does not yet include a way to handle more complex transactions that involve multiple entities paying and receiving different amounts of cash. Such a transaction might look like this:

```
A pays $2
B pays $4
C gets $3
D gets $3
```

Our model requires the addition of a fifth entity to hold contributions before distributions can be made. A fictional or temporary account entity could easily be used for this purpose. Colloquially, we might refer to this account as a "pot" or "kitty," borrowing terminology from card playing games. A different model, one perhaps in which transfers were listed by person or account rather than by event or date, might solve this problem. E.g., the following models the example above.

```
<imaginaryHFR:transaction>
  <imaginaryHFR:person ref="#A"
    gives="$2"
    gets="$0"/>
  <imaginaryHFR:person ref="#B"
    gives="$4"
    gets="$0"/>
  <imaginaryHFR:person ref="#C"
    gives="$0"
    gets="$3"/>
  <imaginaryHFR:person ref="#D"
    gives="$0"
    gets="$3"/>
</imaginaryHFR:transaction>
```

(Note that this example is quite simplified, in that whatever C & D give A & B in return for the \$6 is not included.) However, this sort of model seems overall more cumbersome in the simpler cases. While we note this as an outlier case that our model handles only clumsily, we question how often it would occur in practice, and thus how much of a burden on the encoder of HFRs it would present.

Services

We noted in our poster presented at TEI 2013[3] that services are a problem, but adding the attribute @service to the TEI <measure> element might solve it. We quote from said poster:

Many historical financial records, however, include or are even primarily about the exchange of money for services (e.g., laundering, room and board, or domestic service). Since these services were more usually performed by women and often recorded by women, study of these types of HFRs is of particular interest to practitioners of women's history.

In our “transactionography” we have heretofore used the TEI <measure> element, with its @quantity, @unit, and @commodity attributes, to represent that which is transferred from one person or account to another in a transaction. But in the laundry list case, the work performed by the laundress is not a “commodity” but a “service”, the service for which the boarder paid the boardinghouse keeper in this transaction. However, using the <measure> element with existing attributes leads to markup that fails to distinguish the purchase of a garment from paying for the service of laundering it. One possible solution is to add a new attribute, @service. Thus for instance, a line from a laundry list might be marked up as follows:

```
<hfr:transaction>
    <hfr:transfer fra="#fearn"
                  til="#EW">
        <measure quantity="2"
                  unit="count"
                  commodity="skirt"
                  service="laundering">
            2 wool skirts
        </measure>
    </hfr:transfer>
    <hfr:transfer fra="#EW"
                  til="#fearn">
        <measure quantity="6"
                  unit="pence"
                  commodity="currency">
            6
        </measure>
    </hfr:transfer>
</hfr:transaction>
```

This solution seems to have broad application. E.g.:

```
*Framing:
<measure quantity="27"
          unit="count"
          commodity="8x10"
          service="framing">

*Shoe shining:
<measure quantity="2"
          unit="count"
          commodity="shoe"
          service="shining">

*XSLT programming:
<measure quantity="18"
          unit="hours"
          service="programming">
```

We will not be surprised, however, if there are cases it does not handle well.

Another approach might be to think of “shirt” as the unit and “launder” as the commodity. This would have the advantage of not requiring the addition of an extra attribute to the <measure> element. And since we have marked the unit as “count” in both the examples of framing and shoe shining, these examples might suggest that this latter approach is more elegant than our more verbose use of @service. We wonder, though, whether the programming example might point to a place where the option of greater verbosity constitutes an advantage. Nor does the alternative approach offer a solution to at least two other complexities that occur when we try to formalize references to women’s work in HFRs. We might call these multiplicity and indirection. We quote again from our TEI 2013 poster:

By “multiplicity” we mean that the goods or services being transferred are actually a combination of distinct separate goods or services being transferred as a unit. A common example of this is a suit of men’s clothes, which at one point in the twentieth century might have meant either a jacket and trousers or a jacket, a vest, and trousers.

The problem of “indirection” occurs when the goods or services being transferred are referred to indirectly or metaphorically. ...

These two problems can, and often do, occur simultaneously. For example, the reference to “steak” on a restaurant receipt does not refer to a single chunk of uncooked meat as it would on the receipt from a butcher shop. Rather, it refers to the meat, and perhaps some sauce, in addition to the services of cooking the dish and delivering it to the table.

We do not know, at least not yet, which of these approaches is superior. However, we feel they are probably close enough to each other in expressive power and usability that the important thing, as Tommie Usdin points out[4], is for the HFR encoding community to pick one.

Conclusion

We have considered here some problems that have arisen as we have tested the transactionography model. We are aware of numerous projects in which scholars are working with problems related to markup of HFRs, and we have created a website with space for both discussion of these problems and display of proposed solutions. We invite our colleagues to contribute to Encoding Historical Financial Records <encodinghfrs.org> as we continue to explore TEI-conformant models for markup of these abundant archival records.

Acknowledgement

The authors would like to gratefully acknowledge C. Michael Sperberg-McQueen for not only finding problems with transactionographies, but also allowing us to bounce ideas around.

References

- [1] **Tomasek, Kathryn and Bauman, Syd** (2012), “*Encoding Financial Records for Historical Research*”. Presented at _TEI and the C{rl}O{wu}D: the Text Encoding Initiative Conference, College Station, TX, November 8–10, 2012.
- [2] **Tomasek, Kathryn and Bauman, Syd**, “*Encoding Financial Records for Historical Research*”. *_Journal of the Text Encoding Initiative_*, issue 6.
- [3] **Tomasek, Kathryn and Bauman, Syd** (2013), “*Laundry Lists and Boarding Records: challenges in encoding ‘women’s work’*”. Poster presented at The Linked TEI: Text Encoding in the Web: the TEI Conference, 2013, Rome, Italy, October 2–5, 2013.
- [4] **Usdin, Tommie** (2002), “When “It Doesn’t Matter” means “It Matters””, presented at Extreme Markup Languages 2002, Montréal QC. conferences.idealliance.org/extreme/html/2002/Usdin01/EML2002Usdin01.html

When Kidnapping is but One Risk: Digital Studies Challenge Scholarly and Regional Cultures

Toth, Michael

mbt.rbttoth@gmail.com
RB Toth Associates

Emery, R. Douglas

emery@upenn.edu
Schoenberg Institute for Manuscript Studies, University of Pennsylvania

Contrasting digital studies programs in the United States and Middle East highlight the impact of digitization, open access and digital collaboration on not only users’ cultures in these and other areas, but also the traditional scholarly culture. Early Islamic and Christian manuscripts offer invaluable opportunities to understand and analyze the transfer of mathematics, science and history through ancient texts. Digitization, online scholarship and open sharing offer opportunities for broader access to manuscripts that are at increased risk in some libraries in the Middle East, with free access for citizens and scholars. This open access empowers digital study by scholars and the public in the region and around the globe, while challenging traditional proprietary manuscript research and institutional traditions of protected access and exclusive study. These digital studies also empower communities of the Middle

East – Christian and Islamic – to access their cultural history and artifacts, while highlighting the vulnerability of collections and the need to protect and preserve the digital data.

In 2009, teams of scholars and technical personnel began three major manuscript digital scholarship initiatives:

1. The Walters Art Museum began digitization and online cataloging of Islamic and Western texts for free access, with the support of three successive National Endowment for the Humanities (NEH) grants.¹
2. An integrated team of scientists, engineers, scholars and technical experts used privately funded spectral imaging to reveal the medical undertext of the Syriac Galen Palimpsest for free access in support of ongoing studies.²
3. St. Catherine's Monastery in the Sinai Desert allowed the first large-scale spectral imaging of almost 200 palimpsests from the Monastery library by the Early Manuscripts Electronic Library (EMEL) for scholarly study, under Monastery access controls with support from Arcadia.³

These programs have had to address institutional, cultural and social challenges associated not just with the open study and access to large amounts of data, but amidst cultural upheaval and revolutions. The intellectual property requirements of the owners and institutions also reflect different cultures – from that of a private owner and institution eager to share their collections with the world, to an ancient monastery contemplating how to transition their 1400-year tradition of protecting manuscripts held for centuries within their walls.

Hosting Christian, Islamic and secular texts online highlights the challenges faced in addressing the cultural requirements of different religious and nonreligious traditions – not to mention the challenges of capturing, transferring and accessing data in a hostile region amidst kidnappings and multiple revolutions.⁴ This also poses significant challenges with ongoing global transitions within scholarship and institutions:

- Institutions grappling with shifts from restricted pay-for-access models – based on physical protection of manuscripts – to free-access models with system maintenance costs.
- New generations of scholars capitalizing on digital access with cataloging/data integration tools, while others with limited technical expertise need mediated access with technical support.
- Scholars and institutions struggling with loss of control of their perceived patrimony and/or raison d'être.

The Walters Art Museum in Baltimore, supported by the NEH, is continuing its program to digitize their collection of medieval manuscripts.⁵ This started with digitization of their Islamic Manuscripts and continues by language and manuscript type with their Western Manuscripts.⁶ All data is hosted at thedigitalwalters.org, with new data uploaded regularly.⁷ The data includes “complete sets of high-resolution archival images of manuscripts from the collection of the Walters Art Museum, along with machine-readable TEI P5 descriptions and technical metadata, released for free under a Creative Commons Attribution-ShareAlike 3.0 Unported license⁸ for anyone who wants to use them.” A workflow with virtual cataloging of the manuscripts from the posted images ensures this information is captured in the metadata. Scholars provide cataloging information for each codex and leaf, which is then entered under their name into the digital record of the manuscript. Uses of the data range from the Stanford University Mirador viewer study tool and online Searchworks catalog⁹ to the World Digital Library¹⁰, as well as Flickr¹¹ – where images are used by global members to support a variety of interests. The manuscripts have been downloaded by institutions around the globe, including in the Middle East.

The under text of the Syriac Palimpsest proved to be Galen's medical treatise *On the Mixtures and Powers of Simple Drugs* in a linguistic transition from Greek to Arabic. The bound manuscript is an eleventh-century liturgical text that is also very important for the study of the hymns in its upper text.¹² The private owner made all the image data and metadata available for free access at digitalgalen.net¹³, with license for use under Creative Commons Attribution 3.0 Unported Access Rights.¹⁴ This has enabled transcriptions and studies of the Galen text¹⁵,

, as well as an integrated study of herbs in ancient medicine with analyses of Galen's work in the context of other texts. It served as a catalyst for the “Floriental studies” by a group of technically savvy scholars who are analyzing works ranging from early cuneiform writings to Galen.¹⁶ With this open access, Dr. Grigory Kessel of Philipps-Universität Marburg was able to compare images of the Galen Palimpsest with other Syriac texts and find an additional four missing leaves. Two of these have now been spectrally imaged and are now freely available online^{17 18}, while the search continues for more.

With the open access and collaboration that has developed around humanities data, including from the privately owned Galen Palimpsest and public Walters Art Museum manuscript collection, Archbishop Damianos and the monks of St. Catherine's Monastery are assessing the potential and challenges of digital technology. They currently retain full control of the research data collected by EMEL with UCLA support, and a subset of the data is planned for public web release under the auspices of the Monastery.¹⁹ The Archbishop noted in a project workshop in October 2013: “Digitization offers opportunities for continued preservation of the data on various diverse geographical servers around the globe in the event of a catastrophe.”²⁰ This is highlighted by the political situation in Egypt. The Archbishop also cited the importance of combining digitization and scientific/scholarly study, but with income for the monastery from digital images. To meet the Monastery's wishes, access for initial scholarly study of the data is limited to a team of 20 eminent scholars in the 10 ancient languages found in the palimpsest undertexts, with information shared in an open-source cataloging tool. This changes the study methods of a generation of scholars who have traditionally conducted long-term, in-depth studies of texts with sole access. Currently this requires balancing the requirements for limited scholarly access with those for preservation and accessibility.

Each of these programs is addressing the impact that freely accessible data sharing is having on scholarly study. The integration of technologies and work processes to enhance digital scholarship and collections requires the following capabilities:

- Virtual collaboration with dynamic online cataloging to host and update scholarly findings and shared research with user-friendly tools. Global teams of scholars digitally capture their latest findings and research in standardized catalogs of the Walters' Art Museum and St. Catherine's Monastery manuscripts.
- Broadly accepted standards for integration of data and metadata to ensure digitization of and online access to dispersed collection objects. This supports the digital reunification of diaspora manuscripts, such as the Galen Palimpsest originally from St. Catherine's Monastery.
- Licensed Access to allow appropriate control over global sharing and access, while offering confidence to institutions and scholars that their intellectual property will be protected. This allows the monks of St. Catherine's to share data from manuscripts they have protected for centuries²¹, yet flexibility for institutions to ease into broader access – as demonstrated by the Walters' shift to less restrictive Creative Commons licenses.

Integrated teams of scholars and technical experts ensure scholarly needs are addressed in development of technology and data management. Embedding technically savvy scholars with each of these project teams has ensured analytical and academic research needs are addressed in all project phases, from data and metadata collection through hosting and access to the data.²² Open data pose challenges for an older generation of scholars who have traditionally held onto data until they complete their studies, while empowering a newer generation of scholars capable of collaborating with this data, including in conjunction with other data and tools.²³ Including more technically facile scholars and students on project teams also provides support for the generation of scholars who lack needed technical skills to access and study digital data with appropriate technical tools.²⁴

More important than the cloistered challenges posed by traditional academia are the challenges of preserving the

cultural information and patrimony of communities at risk in the Middle East. These three projects highlight the new opportunities digitization and digital scholarship offer for research and analysis around the globe. This is especially true in the Middle East, where they contribute to understanding of long-standing Christian and Islamic communities and traditions. These traditions and cultures are at risk from secular, religious, economic and social conflict, as are the manuscripts that have supported their development. With data hosted in the United States, the Middle East or elsewhere, free data access and proliferation help ensure preservation of the cultural patrimony, while also supporting communication within and across cultures.

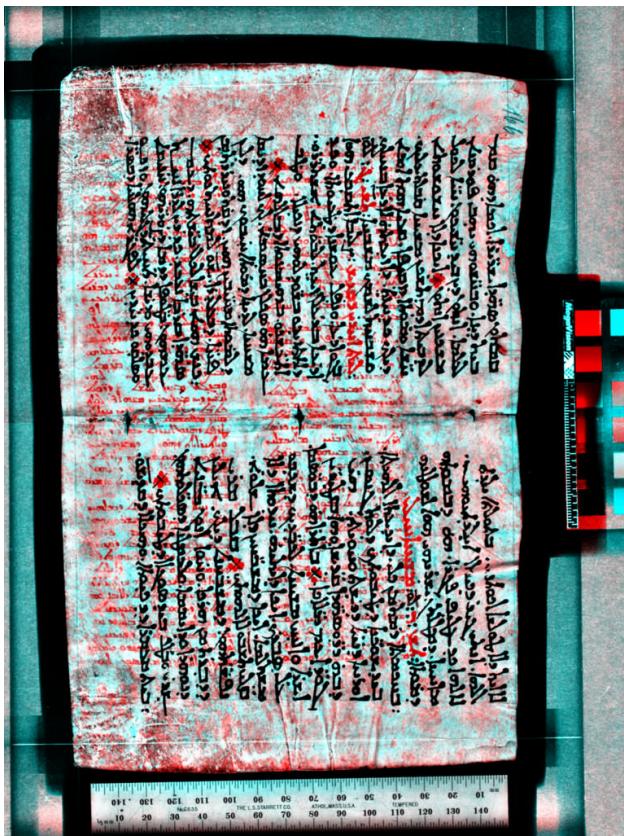


Fig. 1: Galen Syriac Palimpsest Leaf 166r-171v Pseudocolor Image

References

1. D. Emery, M.B. Toth, and W. Noel, *The convergence of information technology and data management for digital imaging in museums*, *Museum Management and Curatorship*, 24: 4, 337 — 356 (2009)
2. S. Bhayro, R. Hawley, G. Kessel, and P. E. Pormann, *Collaborative Research on the Digital Syriac Galen Palimpsest*, *Semitica et Classica* 5 (2012), pp. 261-264
3. Father Justin Sinaites (St. Catherine's Monastery), M. B. Toth , *Spectral Imaging at the Library of St. Catherine's Monastery Reveals Ancient Texts*, Library of Congress, Nov 19, 2012 www.loc.gov/preservation/outreach/tops/ancient_text/ancient_text.html , 30 Oct, 2013
4. M. Shrope, *In the Sinai, a global team is revolutionizing the preservation of ancient manuscripts*, Washington Post, Magazine, www.washingtonpost.com/lifestyle/magazine/in-the-sinai-a-global-team-is-revolutionizing-the-preservation-of-ancient-manuscripts/2012/08/30/1c203ef4-ca1f-11e1-aea8-34e2e47d1571_story.html , 9 September, 2012
5. National Endowment for the Humanities, *Making Medieval Modern - Digitizing medieval manuscripts at the Walters Art Museum*, Division of Preservation and Access Featured Project, www.neh.gov/divisions/preservation/featured-project/making-medieval-modern , 9 April, 2012
6. Walters Art Museum, *The Walters Art Museum Receives \$265,000 NEH Grant to Digitize Over 100 Flemish Manuscripts*,

Press Room, thewalters.org/news/pressdetail.aspx?e_id=365 , 30 Oct, 2013

7. **Walters Art Museum**, *The Digital Walters*, [www.thedigitalwalters.org/](http://thedigitalwalters.org/) , 30 Oct, 2013

8. **Creative Commons**, *Attribution-ShareAlike 3.0 Unported License*, creativecommons.org/licenses/by-sa/3.0/deed.en , 30 Oct, 2013

9. **C. Haven**, *Walters Art Museum manuscript collection makes a virtual move to Stanford*, Stanford Report, news.stanford.edu/news/2013/may/walters-digital-repository-050913.html , 9 May, 2013

10. **World Digital Library**, *Institution: Walters Art Museum*, www.wdl.org/en/search/?institution=walters-art-museum , 30 Oct, 2013

11. **Walters Art Museum**, *Walters Art Museum Illuminated Manuscripts*, Flickr, Photostream, www.flickr.com/photos/medmss/with/5790214510/ , 30 Oct, 2012

12. S. Bhayro, P. E. Pormann, and W. I. Sellers, *Imaging the Syriac Galen Palimpsest: Preliminary Analysis and Future Prospects*, *Semitica et Classica* 6 (2013), pp. 299-302

13. **The Digital Galen Syriac Palimpsest**, digitalgalen.net , 30 Oct, 2013

14. **Creative Commons**, *Attribution 3.0 Unported License*, creativecommons.org/licenses/by/3.0/ , 30 Oct, 2013

15. S. Bhayro and S. Brock, *The Syriac Galen Palimpsest and the Role of Syriac in the Transmission of Greek Medicine in the Orient*, Bulletin of the John Rylands University Library of Manchester 89 Supplement (2012/2013), *Ancient Medical and Healing Systems: Their Legacy to Western Medicine* (ed. R. David), pp. 25-43

16. R. Hawley, *Floriental - From Babylon to Baghdad: Toward a History of the "Herbal" in the Near East*, European Research Council, ERC-2010-StG-263783, 2010

17. **The Digital Galen Syriac Palimpsest**, MS 172, digitalgalen.net/Data/Syriac_MS_172r , 30 Oct, 2013

18. **The Digital Galen Syriac Palimpsest**, SyrNR Frg65, digitalgalen.net/Data/SyrNF-Frg65_001r_50-001/ , 3 Jan, 2014

19. **Universität Wien**, *Sinai Palimpsest Project*:

Alte Schriften neu entdecken, 17 Oct. 2013, medienportal.univie.ac.at/uniview/veranstaltungen/detailansicht/artikel/sinai-palimpsest-project-alte-schriften-neu-entdecken/ , 30 Oct, 2013

20. **Universität Wien**, *The Sinai Palimpsests Project – International Workshop*, 25-27 Oct. 2013, medienportal.univie.ac.at/fileadmin/user_upload/medienportal/uni_view/PDF/PDF-Einladungen_allg/Program.pdf , 30 Oct, 2013

21. **The Digital Galen Syriac Palimpsest**, SyrNF-Frg65 001r, digitalgalen.net/Data/SyrNF-Frg65_001r_50-001/LICENSE.txt , 5 Feb, 2014

22. M. B. Toth, *Integrating Technologies and Work Processes for Effective Digital Imaging*, Proceedings of the Eikonopoia Symposium on Digital imaging of Ancient Textual Heritage, Helsinki, Finland, Oct. 2010, pp. 198-210

23. F.G. France, W. Christens-Barry, M.B. Toth, K. Boydston, *Advanced Image Analysis for the Preservation of Cultural Heritage*, 22nd Annual IS&T/SPIE Symposium on Electronic Imaging, San Jose Convention Center, California, Jan (2010)

24. S. Bhayro, R. Hawley, G. Kessel, and P. E. Pormann, *The Syriac Galen Palimpsest: Progress, Prospects and Problems*, *Journal of Semitic Studies* 58 (2013), pp. 131-148

“Needless To Say”: Articulating Digital Publishing Practices as Strategies of Cultural Empowerment

Tullos, Allen E.

Southern Spaces; Emory Dept. of History; allen.tullos@emory.edu
Emory Center for Digital Scholarship

The open-access, multimedia, peer-reviewed journal *Southern Spaces* will soon begin its tenth year.¹ Published by

the Emory University Libraries, *Southern Spaces* is an online-only journal of critical regional studies that takes as its subject the real and imagined spaces and places of the US South and their global connections. From its years of encouraging digital cultural empowerment as a strategy to transform scholarly publishing, *Southern Spaces* offers a case study for inquiring into the effectiveness of a project intent upon increasing participation in the creation, dissemination, and curation of humanities scholarship. Given the entrenched commercial clout of conventional scholarly publishing and the slow-to-change institutional structures for tenure and promotion that still resist digital humanities scholarship, how can we assess the significance of this project of digital cultural empowerment? And how do the years of experience in publishing *Southern Spaces* contribute to an understanding of broader movements of cultural critique and social change?

Writing about the future of digital scholarly publishing in *Educause Review Online* in 2013, Edward Ayers points to several examples of “acceleration into a full, digital-only environment.”

“Scholars, libraries, and professional organizations in my own field of American history are sustaining innovations in online journals such as *Southern Spaces* and the *Journal of Southern Religion* and in digital meeting places such as *Common-place* and History News Network (HNN). These projects bridge traditional practice and digital possibilities in strategic ways. . . . Blogs and online conversations advance and deepen scholarly conversations, with their impact measured immediately in the number of downloads, views, forwards, comments, and tweets.”

Ultimately, however, Ayers argues that outside of a limited number of examples, “the articles and books that scholars produce today bear little mark of the digital age in which they are created. Thus the foundation of academic life—the scholarship on which everything else is built—remains surprisingly unaltered.”²

Years after the emergence of open-access scholarly publishing platforms, the persistence of institutional inertias, rigid business models, and professional habits of judgment delay and thwart wider deployment of the rich media environment and continually expanding platforms of digital dissemination.³ Established academic print journals have little reason to change their modes of publication and few have done so.⁴ As a result, young scholars, especially in humanities disciplines, find themselves startled by atavistic attitudes. Consider the following excerpt from a letter I recently received from an untenured assistant professor about his journal publishing options:

“I have to share some unfortunate news with you about my contribution to *Southern Spaces*. I’ve recently been able to have a series of conversations with the new director of my program, who is also my advocate and advisor with regard to tenure and promotion—my review will begin next academic year. I discussed *Southern Spaces* with [my director] and while we both think that the online format and the multimedia potential of the format are important to the future of scholarship, her advice to me was that other members of my P/T committee have far more traditional views and would be reluctant to recognize *Southern Spaces* as a publication on par with a print journal. Needless to say, I don’t agree with that assessment, but I think I have to accept it.”

Expressed apologetically, “needless to say” acknowledges the tempting, strong, and rising currents of digital publishing practice that are transforming scholarship by making its production more participatory and “free” (in the sense of free speech),⁵ and its reception widely available. Necessary to say, peer-reviewed *Southern Spaces* essays are used by scholars in a variety of successful professional efforts including tenure and promotion, applications for post-doc fellowships, and landing new kinds of jobs with digital libraries and regional centers.

As part of its intentionality, the collaborative process of producing an open access journal can create a training center where students acquire a range of skills—working with open source software, copyediting, map making, assisting writers and videographers in developing essays, creating technological tools, implementing layout and design—that are valuable in the new digital publishing environment. In a sense, this is a kind of

“building scholarship,” in which “the interventions that occur as a result of building are as interesting as those that are typically established through writing.”⁶ If conceptualized broadly, open-access journal production should also involve a network of independent scholars; researchers associated with nonprofit, grassroots, and nongovernmental organizations; alt-academic writers; as well as the expected cast of college and university-based professors. Supporting this broad understanding of scholarship, *Southern Spaces* (and its social media tools of promotion, e.g., Facebook, Twitter, RSS) has served as an available publishing platform for community-based groups and regional writers engaged in contemporary research related to immigration, environmental destruction, and the crisis of public education. Alerting a networked, topical public about critical writing of interest can produce tens of thousands of readers, as evidenced by a *Southern Spaces* essay on the crisis of government in North Carolina.⁷

With persistent advocacy, the presence of a journal project can lead to institutional commitments for long-term archival preservation of the born-digital materials generated in the work of the journal and into mutually supportive intergroup alliances through organizations such as the recently created Library Publishing Coalition.

It is necessary to say, and to articulate in detail, the various practices of open access publishing that, taken together, comprise a strategy that extends beyond the academy into wider social networks. Drawing upon a decade of publishing practice, my proposed paper will identify and elaborate how a project such as the multimedia journal *Southern Spaces* works to advance cultural empowerment through digital design, creation, dissemination, and curation.

References

1. See southernspaces.org. ISSN 1551-2754.
2. **Edward L. Ayers** (2013), *Does Digital Scholarship Have a Future?* Educause Review Online, August 5 www.educause.edu/ero/article/does-digital-scholarship-have-future?utm_source=Informz&utm_medium=Email+marketing&utm_campaign=EDUCAUSE, accessed October 19, 2013.
3. On the social consequences of habits of judgment see **Judith Butler**, *Giving an Account of Oneself* (New York: Fordham University Press, 2005).
4. For an early statement on the prospects of change that electronic publishing might bring see Bill Kasdorf, “Guest Editor’s Gloss: Reflections on the Revolution,” *Journal of Electronic Publishing*, 3:4 (June 1998). quod.lib.umich.edu/j/jep/3336451.0003.401/-guest-editors-gloss-reflections-on-the-revolution?rgn=main;view=fulltext.
5. *What is free software?* GNU Operating System. www.gnu.org/philosophy/free-sw.html.
6. **Stephen Ramsay and Geoffrey Rockwell** (2012), *Developing Things: Notes toward an Epistemology of Building in the Digital Humanities*, in Matthew K. Gold, ed., *Debates in the Digital Humanities* (Minneapolis: University of Minnesota Press), 83.
7. **Dan T. Carter**, (2013). *North Carolina: A State of Shock*, *Southern Spaces*, September 24. southernspaces.org/2013/north-carolina-state-shock.

Relating texts to 3D-information: A generic software environment for Spatial Humanities

Unold, Martin

martin.unold@fh-mainz.de
i3mainz - Fachhochschule Mainz

Lange, Felix

felix.lange@adwmainz.de
Akademie der Wissenschaften und der Literatur Mainz

1. Introduction

Research in the emergent field of Spatial Humanities, especially in GIS-based approaches, is primarily conducted on a spatial macro-level. Statistical information, such as distributions of archaeological finds or movement and communication patterns is generally mapped to two-dimensional representations of the areas under consideration.¹

²

Humanities research related to spatial contexts is not limited to quantitative scope. Studies in archeology, art history and human geography, to name only a few disciplines, focus on places as small as public spaces and buildings as the stage for liturgical, political and economic practices.³ But important analytical approaches in these fields necessitate three-dimensional representations. Important functional properties of objects in space, such as visibility and accessibility, cannot always be determined by two-dimensional ground plots.

To tackle this shortcoming, the research project Inscriptions in their Spatial Context (IBR) develops research methods and corresponding software tools for Spatial Humanities research on the 3D-geometry of smaller environments like building interiors. Geometric data drawn from geodetic scannings are semantically enriched and connected to scientific textual and visual data. Theoretical foundations as well as the software are being tested in an extensive case study on a medieval church. IBR is a joint research project of the Institute for Spatial Information and Surveying Technology at the FH Mainz - University of Applied Sciences (i3mainz) and the Academy of Sciences and Literature in Mainz. It is funded by the German Federal Ministry of Education and Research (BMBF) and brings together experts from the fields of geoinformatics, surveying engineering, digital humanities, art history and epigraphy.

2. Acquisition of geometries

Panoramic photography has become a common technique for merging smaller photographs into one larger image. Not only established software solutions like Google Street View are able to visualize such panoramic images. Also popular scientific applications⁴ are able to convey the impression of standing at the captured location.

Since images do not represent 3D-information, a measuring campaign is necessary to fulfill the real needs of spatial research. Terrestrial Laser Scanning (TLS) creates high quality point clouds in the range up to several hundred meters.⁵ Until now, tools for processing this kind of information have been mainly designed for the use in the field of engineering. In the Cultural Heritage domain, as in archeology, TLS is usually used for documentation and modeling only or for 2D-analysis, such as for the extraction of floor plans. Parts of the content of the point cloud data are often neither extracted nor used, but can now be reused with the IBR software.

The "GenericViewer" is an easy to handle web-application which provides typical functions of a panorama viewer and allows the user to identify objects in point clouds and to annotate them semantically. Thereby it is possible to work on 3D-objects in an intuitive way, because the software connects panoramic images with point cloud data.⁶

The application communicates with two data storage units: A triple store contains all connections between expert data, spatial data and external resources. Another database handles point clouds, panorama images and spatial objects. It is possible to access the user created data via the GenericViewer itself and through a machine-readable interface. Only Javascript and a browser with WebGL-support are necessary to use the web application.

3. Semantic enrichment

The GenericViewer integrates a customized version of the website-annotator Pundit.⁷ ⁸ Users formulate simple subject-

predicate-object-statements over geometries and texts that are stored in the form of OAC-conformant RDF-triplets. Semantic entities can be drawn from generic ontologies like DBpedia and if more specialized vocabularies⁹ do not suffice, also from project-specific resources.

Physical objects identified in the point cloud are represented by 3D-coordinates, i.e. numerical values. To make them become meaningful objects, some semantic description is necessary. Therefore, geometries can be identified as specimen of a certain type, e.g. a tomb slab, and semantically connected to other textual and geometric resources. The tomb slab, for example, can be tagged as an instance of the class "tomb_slab", and its inscription can be connected to a corresponding critical edition by the relation "is_edition_of". If a researcher wants to suggest a sculptor for this object, knowing that this attribution remains merely a hypothesis and that there might be conflicting opinions, that hypothesis can be expressed as such via suitable predicate. Thanks to the OAC format, annotations come to represent a scholarly discourse rather than just a set of semantic descriptions. The annotation repository can be programmatically accessed and queried via a SPARQL-endpoint, thereby enabling quantitative analyses (e.g. how many tomb-slabs are there in the church) and the creation of aggregate resources with new applications and information services on top of it. The goal is to create a 3D-GIS for quantitative as well as qualitative research with a focus on the interconnection of resources and geometric analysis, as an alternative to approaches centered on multimedia presentations.¹⁰

4. Case study

The GenericViewer is currently being used experimentally in a case study on the late-gothic parish church Liebfrauenkirche in Oberwesel (Middle Rhine Valley, Germany), using 3D-data that has been gathered in a TLS-campaign. The study investigates *inter alia* the placement of tombs and memorial inscriptions. It looks at potentially relevant factors like the social division of the congregation room, procession routes, and other places of liturgical practices. Among the research questions are: What does the placement of a tomb say about the social status of the deceased, given social features of the surrounding tombs? How does an inscription text relate to liturgical texts that were read on procession stations in its proximity? Who could actually see an inscription from which position in the church, and do some texts appear to have been directed at certain groups within the congregation room? The evidence relevant to such questions as well as conclusions drawn from the data are connected to the 3D-representation as semantic annotations.

An important textual source in this study is the epigraphic database "German Inscriptions Online" (DIO), which provides critical editions for the inscriptions under consideration. The spatial configuration of the objects bearing inscriptions and liturgical "key-positions" like the apse and the chancel have been analysed with respect to visibility patterns. The analytical perspective of Space Syntax¹¹, which has been successfully employed in earlier works on historic liturgy and interior church architecture¹², seems to be a promising approach. Due to the complexity of the spatial configurations, our approach offers advantages over conventional text-image representations in terms of analysability and confirmability.

The foundational inscription, a glass painting in the apse of the church¹³, is a case in point. It has been described as a self-assured, even provocative message from the citizenry who build the church to the clergy and especially the bishop. The founders chose the rather exceptional place of the apse windows, because here it was most visible from the choir and the main altar, where the clergy assembled.¹⁴ This presupposition was empirically tested in a visibility analysis.

5. Discussion

At the moment, the GenericViewer is usable with TLS-Data only. However, scanning campaigns can be quite expensive,

and the geometrical accuracy delivered by this technology is not needed in every case. There are other measurement techniques, e.g. "structure from motion", that should be supported in future versions.

Visibility is relative to human perception and it is dependent on a variety of factors, many of which can not currently be modelled in our algorithms. Certain aspects of visibility, e.g. the ability to recognize a certain gesture from a distance, still have to be approximated by human experiments.¹⁵

How useful the data produced with our software is for a broader scientific community will depend on the choice of semantic resource (users are always faced with a trade-off between specificity and generalizability) and the underlying text-document. Pundit, like many similar annotation tools¹⁶, annotates positions in the DOM-tree of HTML-documents. But that means annotating the presentational layer of a text rather than the text itself. Because of that, IBR decided to annotate tree positions in TEI-encoded XML-documents. But not many digital documents in the humanities domain are available in this format, and it is not clear yet how to treat PDF-files.

IBR offers the possibility to contribute to or extend the open source code and to connect third-party applications, analysis tools and ontologies. More informations are available on www.spatialhumanities.de.

References

1. Hernandez, Armando Anaya; Guenter, Stanley P.; Zender, Marc U. (2003): *Sak Tz'i, a Classic Maya Center: A Locational Model Based on GIS and Epigraphy*. In: Latin American Antiquity, Vol. 14, No. 2 (Society for American Archaeology, Jun., 2003), pp. 179-191.
2. v. Lünen, Alexander and Travis, Charles (ed.): *History and GIS. Epistemologies, Considerations and Reflections*. Heidelberg [et al.]: Springer, 2013.
3. Ananieva, Anna; Bauer, Alexander; Leis, Daniel; Morlang-Schardon, Bettina; Steyer, Kristina (ed.): *Räume der Macht. Metamorphosen von Stadt und Garten im Europa der Frühen Neuzeit*. Bielefeld: transcript Verlag, 2013.
4. 3D-viewer of the Romanic dome of Speyer: www.kaiserdom-virtuell.de.
5. Kern, F., Bruhn, K.-Ch., Mehlig, S.: *Messtechnik und Inschriftenforschung - Anwendungsbezogene Arbeiten im Projekt "Deutsche Inschriften Online 3D"*. In: Photogrammetrie, Laserscanning, Optische 3D-Messtechnik - Beiträge der Oldenburger 3D-Tage 2012, hg. v. Th. Luhmann, Ch. Müller, S. 22-33.
6. Kern, F., Mehlig, S., Siegrist, B.: *Geometrische Qualität von aus Einzelphotos zusammengesetzten Panoramen*. In: 29. Wissenschaftliche Jahrestagung der DGPF - Mainz - Geodaten - Eine Ressource des 21. Jahrhunderts (Publikationen der Deutschen Gesellschaft für Photogrammetrie, Fernerkundung und Geoinformation e.V., Band 20, 2011), hg. v. E. Seyfert, S. 129-136.
7. Pundit Semantic Annotation Tool: thelund.it.
8. Grassi, Marco; Morbidoni, Christian; Nucci, Michele; Fonda, Simone; Ledda, Giovanni (2012): *Pundit: Semantically Structured Annotations for Web Contents and Digital Libraries*. In: Mitschick, Annett; Loizides, Fernando; Predoiu, Livia; Nürnberger, Andreas; Ross, Seamus (ed.): *Semantic Digital Archives 2012. Proceedings of the Second International Workshop on Semantic Digital Archives (SDA 2012)*, Paphos, Cyprus, September 27, 2012, CEUR-WS.org/Vol-912.
9. Getty Art & Architecture Thesaurus: www.getty.edu/research/tools/vocabularies/aat/index.html
10. Corrigan, John (2010): *Qualitative GIS and Emergent Semantics*. In: Bodenhamer, David J.; Corrigan, John; Harris, Trevor M. (ed.): *The Spatial Humanities: GIS and the Future of Humanities Scholarship*. Bloomington: Indiana University Press, pp. 76-88.
11. Hillier, Bill: *Space is the Machine: A Configurational Theory of Architecture*, Cambridge: University Press, 1999.
12. Clark, David L. Chatford (2007): *Viewing the Liturgy: A Space Syntax Study of Changing Visibility and Accessibility in the Development of the Byzantine Church in Jordan*. In: World Archeology, Vol. 39 No. 1 (Mar. 2007), pp. 84-104.
13. Epigraphic Database "Deutsche Inschriften Online" (DIO): www.inschriften.net. DIO 60-27: www.inschriften.net/rhein-hunsrück-kreis/inschrift/nr/di060-0027.html.
14. Nikitsch, Eberhard Josef (1996): *Ein Kirchenbau zwischen Bischof und Stadtgemeinde*. Zur angeblich verlorenen Bauschrift von 1308 in der Liebfrauenkirche zu Oberwesel am Rhein. In: JbWdtLg 22 (1996), pp. 95-112.
15. Clark, David L. Chatford (2007): *Viewing the Liturgy: A Space Syntax Study of Changing Visibility and Accessibility in the Development of the Byzantine Church in Jordan*. In: World Archeology, Vol. 39 No. 1 (Mar. 2007), pp. 87.
16. Khalili, Ali ; Auer, Sören; Hladky, Daniel (2012): *The RDFa Content Editor – From WYSIWYG to WYSIWYM*. In: Proceedings of COMPSAC 2012 – Trustworthy Software Systems for the Digital Society, July 16-20, 2012, Izmir, Turkey, 2012.

Dynamic Visualizations In Enriched Publications Of Seventeenth Century Science

Van den Heuvel, Charles

charles.van.den.heuvel@huygens.knaw.nl
Huygens ING (KNAW), The Hague Netherlands

Cocquyt, Tiemen

tiemencocquyt@museumboerhaave.nl
Huygens ING, The Hague/Museum Boerhaave Leiden, Netherlands

Hoogerwerf, Maarten

maarten.hoogerwerf@dans.knaw.nl
DANS (KNAW), The Hague Netherlands

Nagel, Dylan

dylan@wldcrd.com
Wild Card, The Hague Netherlands

Thijssen, Michiel

Thijssen@brill.com
Brill Publishers, Leiden The Netherlands

1. Introduction

Up to now publishers made limited use of the WWW by using websites merely for promotion of their works or as extension of books to accommodate notes, appendices or illustrations that could not be included. This paper describes the results of an interdisciplinary collaboration in which researchers, digital archivists and private companies created web based dynamic visualizations to enrich publications.¹ It presents the outcomes of a pilot project in which Brill Publishers, a game developer Wild Card, historians of science of the Huygens Institute for the History of the Netherlands and the history of science museum Boerhaave and finally scientific programmers of Data Archiving and Network Services (DANS) experimented one day per week for a period of nine months with the development, preservation and reuse of animations of static illustrations of scientific processes and the working of mechanical instruments and tools. These animations were developed to enhance the understanding of complex, often abstract descriptions in seventeenth century publications of the sciences. In total six animations were produced that can be activated by clicking on QR codes next to illustrations in books on/facsimiles of seventeenth century works on astronomy, physics, biology, fortification, land surveying and mechanical engineering. The aim was not just to produce creative illustrations, but interactive scholarly multimedia animations that would contribute to a critical interpretation of the seventeenth century texts and comments hereof. Moreover, the experiment did not limit itself to creation of animated visualizations, but comprehended workflows for re-use and the development of business

models leading to animated discussions about the ownership and responsibility of external links and issues of copy and author rights. The development of criteria for the choice of animations resulted in a typology of enhanced publications. For each of the 6 cases a scenario was developed to model the workflows to enable the publisher and the user to enrich the publications in question. Different types of target audiences were individuated that resulted in discussions of how much freedom of manipulation the user would have and its impact on the storyboard that was developed for each case. After an overview of the implementation of the various storyboards into animations, the architecture of the archive work-flow and business models will be discussed to enable the re-use of the enriched publications in the future.

Toward a Typology of Enhanced Publications

The name of the project that received funding was Dynamic Drawings in Enhanced Publications. DRIVER (2009), the Digital Repository Infrastructure Vision of European Research defines an enhanced publication as one that is enriched with three categories: 1) Research data (evidence of research), 2) Extra materials (to illustrate or clarify) and 3) Post-publication data (commentaries). Our project encompasses enrichments of all these three categories. However, we might have chosen for an alternative definition such as rich internet publication (but we wanted to stress the relationship between the analogue book and the digital enrichment) or scholarly media (but enrichments were not just intended for scholars, but in principle for cultural heritage and education as well). (KAIROS, 2011; Burgess and Hamming, 2011). In hindsight, we probably would have gone for enriched publications, at least if we had followed the distinction that Sondervan (2013) formulated between enhanced publications (hyperlinked content and data with added metadata) and enriched publications that offer more functionality (preferably inside the digital publication itself), with facilities to explore and to analyze data, providing better insight in the underlying research. Instead of contributing to this semantic discussion we created a more pragmatic typology to choose which cases to include for enrichments based on the following criteria: wide coverage of disciplines, degree of interactivity, complexity of dimensions (2D/3D), availability of sources in English and in portfolio Brill publishers and finally the potency for re-use:

Astrolabe: a two-dimensional model of the motions of the heavens and to measure time. The explanation of its working is complex and turned out to be a promising case for exploring and reusing the possibilities of dynamic digital drawings in an interactive way for education, exhibitions and research.

Refraction: Several decades before Newton announced his color theory, René Descartes proposed a physical model for light refraction and to explain the appearance of the rainbow. This case resulted in a discussion about how to visualize and contextualize an “incorrect theory.”

Swammerdam's microscopic drawings: Swammerdam's aim was to highlight the analogous development, in corresponding stages of both the higher and lower animal species. Therefore a visualization with parallel sliders to animate these stages was chosen.

Surveying/Triangulation: This casus zooms in on the Early Modern tradition of surveying treatises, which provide practical instructions for measuring objects in the field and to represent them on drawings. For this casus interactivity in educational context was further explored.

Fortification: Early modern science attempted to tackle real-life problems with applied mathematics. The challenge was here to visualize the mathematical optimization of regular polygonal fortified cities without forcing the user to switch to different levels.

Mill model Ramelli: Agostino Ramelli's *Le diverse et artificiose machine* (1588) was a highlight in the Renaissance Theatrum Machinarum tradition, but its reconstruction raised historic-methodological issues. The question was asked whether additional historical sources could be used for making the mill 'work' with digital tools.

Creating Storyboards

The differences between the six chosen cases implied that for each animation a different storyboard had to be created. However, each story board has the following basic architectural features in common: original source –contextualization (in one or more publication forms) – translation layer and the enrichment. (see Fig.1)

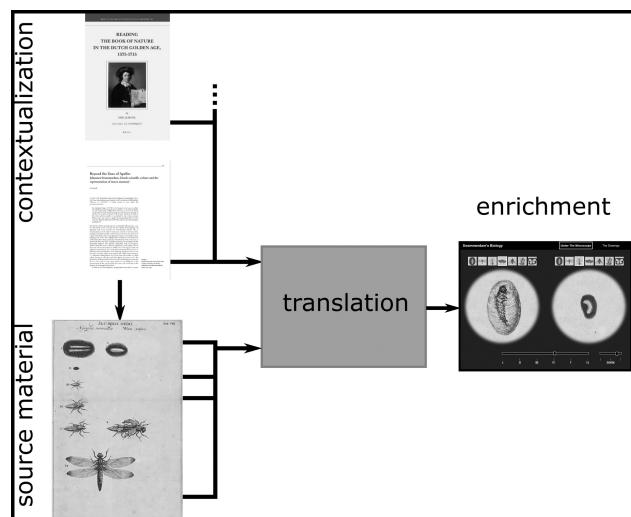


Fig. 1: Storyboard with basic architectural features of enriched publications. Casus Swammerdam

The Astrolabe case

The astrolabe case is not only suitable to illustrate the added value of an interactive animation. In addition to a technical description to explain the working of a complex instruments for research, it is also popular in school and museum environments dealing with education of history of science. Therefore, it has a large re-use potential for non-expert groups as well, which from a business perspective allows spreading the investment over more publication types for different markets. Using the “Integrated Online Exhibit Model” (Marable, 2004) the astrolabe case was translated into a storyboard with three layers: “experience”, “exhibit” and “research” that all could serve as an entry point. The experience layer was used to answer the simple question: “what is an astrolabe?”. In the exhibit layer the user explores interactively functions. Following step by step instructions the user can find the time by measuring and processing solar altitude on the astrolabe (see Fig. 2). The third, “research level” is reserved for all the contextual source material for further in-depth analysis.

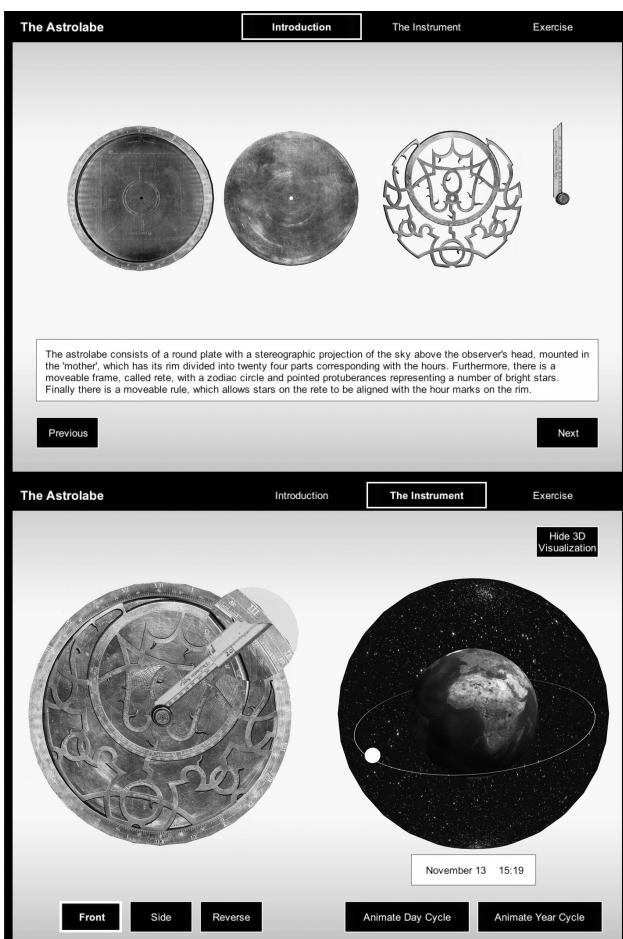


Fig. 2: Experience and Exhibit layer of Enriched Publication. Casus Astrolabe.

Archiving Interactive 3D Visualizations for Re-Use

An archiving workflow was developed with persistent identifiers to ensure the long-term availability of the enriched publication and the reuse of the visualizations. Brill will be the primary publisher for the animated visualizations. DANS will archive the enriched publication, its source files and a durable screencast of the visualization.

Potential Business Models for Publishers

Linked with the Brill content, these enrichments serve as a valuable supplement to current journal articles and books, not only in history of science but in all corners of the humanities & social science spectrum. Open Access publication of the enrichment under CC-BY license seems most suitable. As (governmental) research funding organizations increasingly pose the condition of open access disclosure for research-related data and conclusions, the visualizations should naturally also be freely accessible and 'discoverable' as much as possible; for scholars and non-experts alike. Books will become 'richer' in content and value to the reader. Even if the visualization is available in OA, such books can be priced higher – to recognize the additional value but also to cover costs of creating and facilitating the enrichment. With the outcome of this experiment Brill can ensure the dissemination and preservation of enriched publications. Moreover, the publisher foresees extra services to authors by providing a toolkit for new enrichments or to create those against cost prize, in case an author is unable to do so.

¹ The authors are indebted to Leen Breure (DANS) who provided much information about new publication formats, to Huib Zuidervaart and Eric Jorink (HuygensING) who gave critical feedback respectively on the Descartes and Swammerdam cases and to Valentijn Gilissen and Heleen van de Schraaf (DANS) for their contributions to the development of archiving strategies for enriched publications for reuse.

Financial support we received from the Royal Netherlands Academy of Arts and Sciences and Brill Publishers.

References

- Breure, L., Voorbij, H. and Hoogerwerf, M., (2011). *Rich Internet Publications: "Show What You Tell"*, Journal of Digital Information, 12:1: journals.tdl.org/jodi/index.php/jodi/article/view/1606
- Burgess, H.J. and Hamming, J., (2011). *New Media in the Academy: Labor and the Production of Knowledge in Scholarly Multimedia*. In Digital Humanities Quarterly, 5:3. www.digitalhumanities.org/dhq/vol/5/3/000102/000102.html
- DRIVER (2009). www.driver-repository.eu/Enhanced-Publications.html last updated February 2009. (Accessed March 2014)
- KAIROS (2011). pkp.sfu.ca/ocs/pkp/index.php/pkp2011/pkp2011/rt/metadata/287/0 . (accessed March 2014)
- Marable, B., (2004). *Experience, Learning, and research: Coordinating the Multiple Roles of on-line Exhibitions, Museum and the Web*, 2004: www.archimuse.com/mw2004/papers/marable/marable.html .
- Sondervan, J., (2014). 'What stakeholders think of enhanced publications.' Blog post April 10th, 2013 www.sondervanpublishing.nl/?cat=5 (Accessed March 2014).

Process Data for Digital Scholarly Editions

Vasold, Gunter

gunter.vasold@uni-graz.at
University of Graz

Introduction

Working on scholarly edition in general can be seen as a chain of working steps. Each step leads to an intermediate result, which other steps may build upon. At the end of this chain stands the published edition as the final result. Traditionally these steps, thus the editorial process, are described in form of methodological guidelines and editorial principles. Digital scholarly editions not only add new aspects like the separation of data and presentation or the chance to include additional materials ^{1 2 3 4}, they might also change the way we document editorial processes. This seems particularly important for new forms of digital scholarly editions like *open source editions* ⁵, *work in progress editions* ⁶, *editions 2.0* ⁷ or *social editions* ⁸. Such editions are evolving over time, they are possibly incomplete (work in progress), they might have been created in a collaborative way by many editors, or they provide not only canonical texts but also accompanying materials or results from intermediate stages like graphemic transcriptions, pre-normalized texts etc. as raw data for further enrichment and research.

As a consequence it is questionable if traditional ways of documenting how an edition was made are still appropriate. If we want to deal with the fact that digital scholarly editions are no longer static resources but collections of evolving results, we must put a stronger focus on process aspects. For evolving scholarly editions we should consider to understand formal process documentation as an essential part of such editions.

Problems

First we have to define what we mean with editorial processes. Creating a scholarly edition includes a large number of activities like searching, collecting and evaluating sources, other textual witnesses, literature and images, the review of existing and preliminary editorial work, making transcriptions,

collations, normalized texts, abstracts, indexes, glossaries. It also implies tasks like dicscrimen veri ac falsi, stylometric, paleographic, prosopographic, and sphragistic research. Some of these tasks have to be done only once, but the majority of editorial activities is either recursive and/or has to be repeated over and over again, e. g. for each charter of a collection.

Documenting such processes seems easy, but if we take a closer look, we find that the documenting of single processes requires a clear and formalized concept of processing steps based on the definition that a process is a directed activity which leads from one state to another. I assume that editorial activities can be separated from each other and can therefore be handled as addressable and formally describable units of action. This is a basic prerequisite for process documentation, because we need distinct entities to refer to. Each process has to be defined with its starting point, a defined ending point and a set of activities which lie in between. This means that we do not only have to describe activities, but we also have to create relations between a specific activity and the data sets which represent the starting and ending points of the activity.

Another issue is how we should describe editorial processes. This includes three basically inseparable sub-questions: (1) what should be documented to what extent (2) how should this be put into practice and (3) which data formats are feasible. The extent of documentation depends on the objectives of an editorial project, but a minimal set should include timestamps, actors (persons or algorithms), a formalized description and possibly an additional free-text description of the action and, as mentioned before, pointers to the initial and resulting data.

The second question deals with the problem that documenting single editorial activities seems to put a massive overhead to editorial work. But for the greater part this data can be generated automatically if we use appropriate working environments. The question of feasible data formats is twofold: we have to distinguish between process data used within a project and data which becomes a formal part of the edition. The former needs little discussion, as it depends on technical environments. But the latter is still an open problem which requires further discussion and research although some solutions already exist. It seems that graph based models are particularly suitable for the problem^{9 10}, but also XML-based standards for process descriptions like PREMIS must be investigated as possibly adaptable models.

The third and most fundamental question is what can we do with this data or more pragmatic: why should we start collecting process data at all? Formalizing editorial processes and the availability of such data has multiple benefits. From a project management point of view this data can be used to get an overview of the project status at any point in time. One can use this data to find out for example how long a task takes on average, which might become important for future project planning. Process data also can be used to trigger actions like sending a message to co-workers if a certain state has been reached or even to move a document automatically along a predefined workflow based chain of tasks. In quality control process data can be utilized to identify systematic errors.

From a user point of view quality control will become more important with new forms of scholarly editions. If an edition has many contributors, or if an edition has no final result, it is partly up to the user to decide if a particular element of an edition is trustworthy. Peter Robinson has claimed that every act of editing should be attributed to the person who did it, because this allows attribution and trust¹¹. Process metadata describing when and by whom an element was added or changed can therefore be an important indicator for the reliability.

Open ended and collaborative editions consist of an ever increasing number of representation forms and versions, which constitute multidimensional knowledge spaces. Process data can be used to construct references and therefore contexts between the single components of such a multidimensional scholarly edition¹². Hence, process data can give important directions when users want to leave the paths predetermined by editors.

Approaches

The problem of process planning and process documentation is of course not unique to scholarly editions. One example is software development, where many contributors can work on the same project, continuously creating new and refactoring existing code. This requires the use of software versioning and control system like Apache Subversion or Git for the creation and organization of process metadata which document different versions of files and development branches. In science there is a long tradition of hand written log books with the purpose to keep laboratory work reproducible and transparent. Nowadays these log books are replaced by electronic systems which generate and log great parts of process metadata automatically¹³.

Most advanced in this domain is business process management (BPM) as a discipline of business informatics¹⁴. It is focusing on process optimization which is not the main problem with scholarly editions, but BPM has developed a domain specific terminology, conceptual frameworks, software tools and standards like WfMC, WSFL, XPDL, BPEL or BPMN, which turn out to be very useful for understanding our topic. Concerning the problems they try to solve document related technologies¹⁵, especially (enterprise) content management^{16 17} seem to be more applicable to digital scholarly editions than BPM. Content management describes the life cycle of a document through a chain of states in an abstract way. Content management systems implement these formalized life cycles by providing states, workflows, users and roles. They are able to keep track of stages and versions, are able to automatically route documents through defined paths and to collect metadata for every stage and action. Enterprise content management must not be confused with web content management, as web content management systems are mostly simplistic, monolithic and proprietary applications, which are not suitable in terms of long time availability.

Conclusion

New open forms of digital editions require a stronger focus on editorial processes. This affects the planning of work, needs support for process specific workflows and requires process specific metadata which has to be seen as an essential part of an edition. Other fields, particularly (enterprise) content management has requirements similar to those of scholarly editions. Upcoming software tools for digital editions should investigate CMS with regard to process aspects and adopt useful features. Special focus has to be put on the development of an appropriate, interchangeable format for edition-related process metadata.

References

1. Sahle, P. (2013). *Digitale Editionsformen, Vol. 2: Befunde, Theorie und Methodik*. BoD: Norderstedt.
2. Robinson, P. (2013). *Towards a Theory of Digital Editions*. In: Variants – Journal of the European Society for Textual Scholarship, 10, 105-131.
3. Pierazzo, E (2011). *A rationale of digital documentary editions*. In: Literary and Linguistic Computing, 26, 463-477.
4. Dahlström, M. (2004). *How Reproductive is a Scholarly Edition?* In: Literary and Linguistic Computing, 19, 17-33.
5. Bodard, G. and Garcés, J. (2009). *Open Source Critical Editions: A Rationale*. In: Deegan, M., Sutherland K. (eds.) *Text Editing, Print and the Digital World*. Farnham: Ashgate, 83-98.
6. Kropač, I. H. (2008). *Work in Progress: Vom Digitalisat zum edierten Text*. In: Thumser, M., Tandecki, J. (eds.) *Editionswissenschaftliche Kolloquien 2005/2007*. Toruń: TNT, 167-183.
7. Boot, P. and van Zundert, J. (2011). *The Digital Edition 2.0 and The Digital Library: Services, not Resources*. In: *Bibliothek und Wissenschaft*, 44, 141-152.
8. Osborne, R. et al. (2013). *eResearch Tools to Support the Collaborative Authoring and Management of Electronic Scholarly Editions*. In: Digital Humanities 2013, Conference Abstracts, 334-337.

9. Osborne, R. et al. (2013). *eResearch Tools to Support the Collaborative Authoring and Management of Electronic Scholarly Editions*. In: Digital Humanities 2013, Conference Abstracts, 334-337.
10. DeFerrari, J. et al. (2002). *Managing the Lifecycle of Information*. Hamburg: AIIM.
11. Robinson, P (2013). *Five desiderata for scholarly editions in digital form*. In: Digital Humanities 2013: Conference Abstracts, 355-356.
12. Vasold, G. (2014). *Progressive Editionen als multidimensionale Informationsräume*. In: Ambrosio, S. et al. (eds.) *Digital Diplomatics. The Computer as a Tool for the Diplomatist?* AFD Beihet 14, Köln: Böhlau, 75-88.
13. Potthoff, J. et al. (2011). *Elektronisches Laborbuch: Beweiswerterhaltung und Langzeitarchivierung in der Forschung*. In: Schomburg, S. et al. (eds.) *Digitale Wissenschaft*. Köln: hbz, 149-156.
14. Vom Brocke, J. and Rosemann, M. (2010). *Handbook on Business Process Management*. Springer: Berlin.
15. Kampffmeyer U. (2003). *Dokumenten-Technologien*. Hamburg: Project Consult.
16. DeFerrari, J. et al. (2002). *Managing the Lifecycle of Information*. Hamburg: AIIM.
17. Cameron, S. (2011). *Enterprise Content Management*. Swindon: BCS.

The opportunistic librarian: A Leuven confession

Verbeke, Demmy

demmy.verbeke@arts.kuleuven.be
KU Leuven

Quite a lot of ink has been spilled, or, at least, quite a lot of keys have been hammered on the definition and the history of Digital Humanities. This has also led to a, sometimes heated, debate on “who’s in” and “who’s out”. Regardless of whether all parties concerned are entitled to call themselves “digital humanists”, the consensus nevertheless seems to be that Digital Humanities are in essence a collaborative activity (see, e.g., Siemens 2009¹ and Spiro 2013²), involving academic staff, students, computer programmers, librarians, project coordinators, administrative staff, and others. It is clear that academic librarians should get involved, not only because they are “as much a part of the (Digital Humanities) plan as faculty are” (Pannapacker 2012³), but also because they should take to heart the goals for practitioners of Digital Humanities as they were identified by Lisa Spiro (Spiro 2011⁴; see also Vandegrift and Varner 2013⁵), namely: to provide wide access to cultural information, to enhance teaching and learning, to transform scholarly communication, to enable the manipulation of data, and to make a public impact. Even in a minimal, unambitious definition of the mission of any research library, several of the named goals should form its core business, its heart and soul, and its reason for being. And although the ways in which research libraries have translated these goals into actual activities may have changed considerably during the past decades, the essence still remains the same: at the very least, these libraries aim to provide access to cultural and scholarly information and to enhance teaching and learning at universities and beyond; and are condemned to either continue to do so in the digital age or to become obsolete (Verbeke 2013⁶).

Naturally, the exact nature and form of the contribution of the university library and its staff to Digital Humanities projects at a particular academic institution will differ, depending on who else is already involved, on the available staff and infrastructure, on the wishes and willingness of faculty and administration, and on the ambitions and priorities of the institution in question. Possibilities range anywhere from providing basic information about existing tools for Digital Humanities research and teaching as well as the organization of training sessions so that academic staff and students learn how to use these tools, to the creation of library-based skunkworks – or semi-independent,

research-oriented software prototyping and makerspace labs (see, especially, Nowviskie 2013⁷). Whatever the exact form, however, it seems natural for research libraries and their information professionals to focus on a contribution which is identical or very similar to the roles which are traditionally expected of them anyway, such as discovery and dissemination of data and knowledge, data management, digitization and preservation (Showers 2012⁸ and Vandegrift 2012⁹). Of the latter, William Kretzschmar and William Gray Potter even stated – not without a sense for drama – that “collaboration with the university library is the only realistic option for long-term sustainability of digital humanities projects in the current environment. ... If digital humanities projects stand still, they will indeed die, and the library is the only part of our institutional structure that can keep them moving enough to save them” (Kretzschmar and Gray Potter 2010¹⁰).

This short paper not only evaluates the recent scholarship on the role of libraries and their staff in Digital Humanities projects, but also documents the efforts of the University Library at KU Leuven (Belgium) in general, and the Arts Faculty Library in particular, to maintain, in a Digital Humanities context, its role as an important partner in research. It discusses the various initiatives to transform the ways in which the libraries concerned have supported learning and teaching for decades, and presents various projects to which staff members of the library contribute. Examples include efforts to develop OCR applications for early printed Dutch and Latin texts, the creation of virtual research communities for international projects involving KU Leuven faculty, the building of an integrated reference database and collaborative platform for the study of Patristic, Medieval and Byzantine texts, and a strong involvement in *Europeana* and Linked Data initiatives. Finally, this short paper also briefly presents the current state of the plans to found a Digital Humanities Library Lab @ Leuven (DH3L), offering a frank discussion of the common challenges (see, especially, Posner 2013¹¹) encountered over the past months and the ways in which faculty, administration and library staff at KU Leuven have tried to overcome them.

References

1. Siemens, Lynne. (2009). *It's a Team If You Use 'Reply All': An Exploration of Research Teams in Digital Humanities Environments*. *Literary and Linguistic Computing* 24 (2): 225-233. doi:10.1093/linc/fqp009.
2. Spiro, Lisa. (2013). *Group and Method: Collaboration in the Digital Humanities*. digitalscholarship.wordpress.com/2013/04/10/group-and-method-collaboration-in-the-digital-humanities/.
3. Pannapacker, William. (2012). *No DH, No Interview*. The Chronicle of Higher Education, July 22. chronicle.com/article/No-DH-No-Interview/132959/.
4. Spiro, Lisa. (2011). *Why the Digital Humanities?* digitalscholarship.files.wordpress.com/2011/10/dhglca-5.pdf.
5. Vandegrift, Micah, and Stewart Varner. (2013). *Evolving in Common: Creating Mutually Supportive Relationships Between Libraries and the Digital Humanities*. *Journal of Library Administration* 53 (1): 67-78. doi:10.1080/01930826.2013.756699.
6. Verbeke, Demmy. (2013). *Digital Humanities En de Wetenschappelijke Bibliotheek van de Toekomst*. *Bladen Voor Documentatie / Cahiers de La Documentation* 67 (1): 13-16.
7. Nowviskie, Bethany. (2013). *Skunks in the Library: A Path to Production for Scholarly R&D*. *Journal of Library Administration* 53 (1): 53-66. doi:10.1080/01930826.2013.756698.
8. Showers, Ben. (2012). *Does the Library Have a Role to Play in the Digital Humanities?* infteam.jiscinvolve.org/wp/2012/02/23/does-the-library-have-a-role-to-play-in-the-digital-humanities/.
9. Vandegrift, Micah. (2012). *What Is Digital Humanities and What's It Doing in the Library?* In *The Library with the Lead Pipe*. www.inthelibrarywiththeleadpipe.org/2012/dhandthelib/.
10. Kretzschmar, William A., and William Gray Potter. 2010. "Library Collaboration with Large Digital Humanities Projects."

Literary and Linguistic Computing 25 (4): 439–445. doi:10.1093/linc/fqq022.

11. Posner, Miriam. 2013. "No Half Measures: Overcoming Common Challenges to Doing Digital Humanities in the Library." Journal of Library Administration 53 (1): 43–52. doi:10.1080/01930826.2013.756694.

Kinematics: big cultural data and the study of cinema

Verhoeven, Deb

deb.verhoeven@deakin.edu.au
Deakin University

Coate, Bronwyn

bronwyn.coate@deakin.edu.au
Deakin University

Arrowsmith, Colin

colin.arrowsmith@rmit.edu.au
RMIT University

Davidson, Alwyn

alwyn.davidson@deakin.edu.au
Deakin University

Introduction

The Kinematics Project is a multidisciplinary study of the industrial geometry of culture focussing in particular (but not exclusively) on the cinema. The project results from both the recent digitisation of the cinema industries as well as contemporary research practices in the discipline¹. The researchers have collated a unique dataset of global cinema showtimes which alone, and in combination with additional datasets, challenge many of the unspoken assumptions and ordinary practices of conventional film studies research.

The Kinematics Project proceeds from the emergent understanding that the cinema is not an isolated set of practices. Cinema comprises institutional, social, and commercial networks that are interdependent which in turn influence and shape our approach to cinema research. This view of cinema is relatively recent. To date, the study of cinema has been predominantly concerned with issues of film content (the text) and with little regard for the events that occur around the actual consumption of film². By shifting the focus from film content to cinema as a cultural practice we open the way for new questions and approaches to research that effectively draws together a number of discipline areas. This also distinguishes our work from others with a more formally or textually focussed approach to the computational turn in Cinema Studies such as Lev Manovich's³ pioneering studies. In particular the advent of big data has meant that a wider range of digital data types, formats, and sources can be used in innovative ways by all disciplines including the humanities and social sciences. The availability of big cultural data enables the unprecedented mapping of the industrial geometry of motion pictures at an international scale. This paper uses three case studies to demonstrate how the digitisation of cinema can be understood as a set of located and network practices.

A Big Cultural Dataset to Track Film Flow and Diffusion Across the Globe

Over a 12 month period, we have tracked the global flow of film screenings by gathering specific cinema location information for over 47,000 films throughout 48 countries internationally. For each of these 48 countries we have data for every film screening event (down to date and time for each screen) for all venues (a global total of 30,000) resulting in a database of over 120 million records. Data was obtained from a third party source and is directly downloaded to the

project database. Their data comes directly from cinema venues mostly through automated electronic means and also email and phone calls. Until now, databases dealing with cinema consumption and exhibition have been limited to case studies that are either national scope or defined by special interests. Examples include, historical database initiatives such as the substantial Dutch database "Cinema in Context"⁴ and the "Cinema and Audiences in Australia Project" (CAARP) database⁵ as well as the GIS based work of Robert C. Allen⁶ at the University of North Carolina and the "Australian Cinemas Map" database in Australia⁷. This project extends the scope of such databases by taking it to an international scale, to create the first global study of the film industry.

Method and Analytical Approach

In this paper we will demonstrate how we can further our understanding of cinema as a set of network practices both economically and geographically through the collection of digital datasets and utilising new technologies for analysis. This will be addressed in three linked projects, all of which focus on the global reach of cinematic data and practices. Whilst there are some differences in method across the projects, each of the inter-related projects feature the use of visualisation to explore, analyse, and communicate the information and findings. Visualisations are used throughout the process as it is an effective way of dealing with big data, making the proliferation of data readily accessible.

Each of the related projects are briefly summarised as follows:

1) Tracking the global movement of films

The success of the film *The Hobbit: An Unexpected Journey* has been taken as a case study to track the movement of film at an international scale. *The Hobbit* was chosen due to the challenges it poses from the large amount of viewings, the geographical reach of its screenings, and also the complex temporal and spatial itineraries involved in 'staggered' film releasing strategies. The use of GIS and temporally sensitive visualisations have enabled us to track the spatial and temporal relationships of *The Hobbit*, highlighting the complexities of international cinema enterprises and the subtleties of contemporary releasing strategies.

2) Interoperating data – linking remittances and the movement of film at the cinema

Using India as a case study we explore the relationship between remittance flows and the movement of film around the globe. India provides an ideal case in point to study this relationship given that a relatively large proportion of Indians work abroad sending remittances back home and that India has its own unique highly successful global film industry emanating from Bollywood. By merging data on the flows of remittances from countries returning funds to India with data on the screening of Indian film we have used visualisation and economic modelling techniques to identify the pattern of movement of Indian film around the world, with particular focus on how Indian film flows into the various remittance-sending countries. This analysis is based on bi-lateral remittance flow data sourced from the World Bank. In order to investigate the importance of remittances as a factor helping to explain the flow of Indian film around the world we scale remittances relative to the size of the sending countries overall population. As a result, we are able to test whether a greater presence of Indian nationals within a given country supports a higher level of cultural diffusion.

3) Spatial and temporal persistence in distribution patterns during a period of industrial transition

The distribution industry has been explored through modelling the spatial and temporal attributes associated with the diffusion of films. Although the digitisation of the film itself has opened up the possibilities of new distribution markets and strategies, we have found that there is still a strong relationship to pre-digital distribution territories. Through a number of visualisation techniques including Network Analysis and Circular Statistics, our analysis has also found that there is great variation in distribution patterns dependent on variables such as genre, production company, and country of origin.

Conclusion

Rather than measuring the comparative cultural value of film texts as favoured by traditional cinema studies, the Kinomatics Project traces the flows and pace of industrial change in the cinema and measures the intensity of its dynamics. This paper describes the intersection between a revised qualitative cinema historiography (focused around an industrially informed and consumption attentive view of the cinema) and the use of innovative information systems inspired by new research approaches found in big data analytics such as data mining and digital visualisations.

References

1. Verhoeven, D. (2012), *New Cinema History and the Computational Turn*, in Beyond Art, Beyond Humanities, Beyond Technology: A New Creativity, Proceedings of the World Congress of Communication and the Arts Conference, University of Minho, Portugal
2. Bowles, K and Maltby, R. (2009), *What's new about New Cinema History?*, in H. Radner, P. Fossen (eds), Remapping Cinema, Remaking History: XIVth Biennial Conference of the Film and History Association of Australia and New Zealand, Centre for Research on National Identity, University of Otago, NZ, Dunedin, New Zealand, pp. 7-21
3. Manovich, L. (2001), *The Language of New Media*. MIT Press, Cambridge, MA
4. Dibbets, K. (ed) (2013), *Cinema Contexts*, www.cinemacontext.nl (accessed 12 February 2014)
5. Verhoeven, D. (2013), CAARP, caarp.edu.au (accessed 12 February 2014)
6. Allen, R. (ed), (2013), *Going to the Movies*, docsouth.unc.edu/gtts/ (accessed 12 February 2014)
7. Maltby, R and Walsh, M. (ed), (2013), *Australian Cinemas Map*, auscinemas.flinders.edu.au/ (accessed 12 February 2014)

Less explored multilingual issues in the automatic processing of historical texts – a case study

Vertan, Cristina

cristina.vertan@uni-hamburg.de
University of Hamburg, Germany

1. Introduction

Recently, the collaboration between the Language Technology community and the specialists in various areas of the Humanities has become more efficient and fruitful due to the common aim of exploring and preserving cultural heritage data. It is worth mentioning the efforts made during the digitisation campaigns in the last years and within a series of initiatives in the Digital Humanities, especially in making Old Manuscripts available in the form of Digital Libraries.

Having in mind the number of contemporary languages and their historical variants, it is practically impossible to develop brand new language resources and tools for processing

older texts. Therefore, the real challenge is to adapt existing language resources and tools, as well as to provide (where necessary) training material in the form of corpora or lexicons for a certain period of time in history.

Another issue regarding historical documents is their usage after they are stored in digital libraries. Historical documents are not only browsed but together with adequate tools they may serve as basis for reinterpretation of historical facts, discovery of new connections, causal relations between events etc. In order to be able to make such analysis, historical documents should be linked among themselves and should be linked with modern knowledge bases. Activities in the area of Linked Open Data (LOD) play a major role in this respect

Most digital libraries are made available not only to researchers in a certain Humanities domain (e.g. classical philologists, historians, historical linguists), but also to common users. This fact has posed new requirements to the functionalities offered by the Digital Libraries, and thus imposed the usage of methods from Language Technology for content analysis and content presentation in a form understandable to the end user.

There are several challenges related to the above mentioned issues:

- Lack of adequate training material for real-size applications: although the Digital Libraries usually cover a large number of documents, it is difficult to collect a statistically significant corpus for a period of time in which the language remained unchanged.
- Historical variants of languages lack firmly established syntactic or morphological structures thus the definition of a robust set of rules is very difficult. Historical texts often constitute a mixture of multilingual paragraphs including Latin, Ancient Greek, Slavonic, etc.
- Historical texts contain a large number of non-standardized abbreviations.
- The conception of the world is somewhat different from ours, which makes it more difficult to build the necessary knowledge bases.

Whilst these issues are generally accepted there is still less research done towards the content annotation of old texts. Language technology tools are mostly adapted in order to make corpus linguistics research (Piotrowski 2012), (Vertan et. al 2012) but not really used as means for text processing. One of the main barriers is the text readability in terms of language style and terms.

The aim of this paper, describing on-going work is to demonstrate that multilingual aspects in historical texts are a big challenge but can serve also for building a knowledge-based to be used in text presentation.

2. Selected materials

The selected texts are works of Dimitrie Cantemir, prince of Moldavia at the end of the XVII century, but also historian, philosopher, composer, musicologist, linguist and much other.

As member of the Prussian Academy of Sciences he was asked to write a history of the Ottoman Empire as well as a history of Moldavia, both unknown territories for Western Europe at this time. These remarkable works remained until the middle of XIXth century the only generally accepted reference. Even if some historical aspects are interpreted in a subjective manner, the works of Cantemir represent a unique testimony of that time. He is not describing only dates and places of historical events but presents daily life, occupations, country organisation as he saw himself as prisoner in Istanbul and prince of Moldavia.

The „History of growth and decay of the Ottoman Empire“ and the §Description of Moldavia“ were written initially in Latin. Later they were translated in German, and then the German Edition was translated in English, French, Romanian and Russian. All these Editions are no tone-to-one translation but in many cases they are influenced by the perception of the translator.

We consider that the works of Cantemir are of particular importance for the history of Eastern Europe, about which with exception of specialists, is less known.

The current project aims at the presentation and explanation of these works, by means of language technology tools. For the moment we concentrate on the German, English and Romanian Editions of the „Description of Moldavia“, all available in the Library of our institution and intend to extend the work to the other language editions.

3. Exploiting Multilinguality

Usually the multilingual problem of the old texts is resumed to the fact that Latin and sometimes ancient Greek passages are found. The works of Cantemir have the particularity that they introduce words in the language of the described country, namely Romanian and Ottoman Turkish.

Therefore a processing step is needed in order to identify these words. State-of-the-art methods in language identifications based on n-grams cannot be used as both old Romanian were written with other alphabets (church Slavonic) and the transliteration rules were not standardised.

Our method uses the multilingual versions of his works help here in identifying the foreign words.

Explanation paragraphs differ slowly but the terms are preserved. Therefore we use comparison at sentence level in order to identify common words, which we mark afterwards as Named Entities.

Following example from the „Descriptio of Moldavia“ in German is illustrative:

„Der Watawul de Aprodsci de Tyrg ist Herr über die Gerichtsdienner, welche den Tribut und andere Abgaben der Bürger eintreiben, und andei Schatzkammer liefern. „

Here we observe the following: the expression „Watawul de Aprodsci de Tyrg“ is for a language technology tool dealing with German completely noisy. On the other hand the expression is an approximate transliteration of old Romanian written in Church Slavonic. Nowadays the correct transliteration would be „vatavul de aprozi de târg“.

In comparison here the entry in the R Romanian version

„**Vatavul de aprozi de târg; este asupra slujitorilor divanului, cari strâng birul si alte dări ale orăsenilor si le aduc la Vistrie;**“

Thus the algorithm for the processing of the text involve the following steps:

1. Sentence splitting of German and English and Romanian texts
2. Sentence alignment by means of comparable corpora sentence extraction (Smith et. al. 2010)
3. Identification of common identical strings
4. Extraction of common passages and normalisation using resources in old Romanian / ottoman Turkish as well as heuristic rules
5. Marking of identified expressions as named entities in order to be blocked from further processing.

We use the output of this process in order to built a multilingual knowledge base, and match the terms on wikipedia entries.

4. Conclusions and further work

Through this contribution we want to raise the attention on less studied aspects of multilinguality in historical texts and show how methods from multilingual language technology, here exploitation of comparable corpora, can be used for dealing with this issue. In the presentation we would describe in detail the algorithms used for the construction of the multilingual knowledge base and give relevant examples.

As mentioned before, this is an on-going project. After creating the knowledge base we intend to built a presentation interface where entries in the knowledge base may-be available at the moment of text reading through a mouse-over function. We intend also to asses how much from the found entries match Wikipedia, and show that in fact the information is

complementary. We intend also to extend the methods to other works.

References

Cantemir, C., *Descriptio Moldaviae Historisch-geographisch und politische Beschreibung der Moldau, nebst dem Leben des Verfassers (lat. - 1714-1716); germ. Beschreibung der Moldau*, A. V. Büsch (Ed.), 1770

Cantemir, C., *Geschichte des Osmanischen Reichs nach seinem Anwachsen und Abnehmen. Beschrieben von Demetrie Kantemir, ehemaligem Fürsten in Moldau. Nebst den Bildern der Türkischen Kaiser*, Hamburg Herold, 1745.

Piotrowki, M., *Natural Language Processing for Historical Texts*, Synthesis Lecture on Human Language Technologies, 2012

Smith, J and Quirk, C. and Toutanova, K., *Extracting parallel sentences from comparable corpora using document level alignment*, Proceedings of the HLT '10 Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pages 403-411

Vertan, C. and Osenova, P. and Slavcheva, M. and Piperidis, S., *Adaption of Language Processing and Tools for cultural heritage*, Proceedings of the Workshop at LREC 2012.

Modelling digital edition of medieval and early modern accounting documents

Vogeler, Georg

georg.vogeler@uni-graz.at

Zentrum für Informationsmodellierung - Austrian Centre for Digital Humanities, Universität Graz

Accounting documents seem to be well suited digital reproduction: they contain a significant number of highly structured information which was meant to be calculated already in its original context. Creating a database for executing statistical and numerical analysis of the data seems to be a promising method to analyse these documents. This assumption is not reflected in the practice of critical scholarly editions of accounting books. The digital humanists creating resources of accounting documents have not succeeded to build a common form of representation apart from the preference of spread-sheet style data storage. This approach neglects important information in the original documents which become in particular clear when studying medieval account books:

1. The recent scholarly edition of the Luxemburg city accounts¹ shows that accounting is close to everyday language and thus an important source for research on the history of a language. A digital resource concentrating on numbers, short descriptions of bookkeeping entries and classifications excludes that information.
2. Producing an account in the middle ages was a process which included several steps: collecting the information from informal documents into a fair copy, reading the listed transactions to a supervising committee or to a person using the abacus, deleting debit entries with their fulfilment, entering debits for tax or rent collection and updating the entries during money collection while tax collection or in manorial administration. Medieval accounts thus can be a protocol of acts of accounting and controlling.
3. The layout of medieval accounts changed in time, i.e. for example in administrative accounting the layout developed from protocols in simple text blocks to sophisticated tabular representations². The early forms of double entry bookkeeping are based on the position of the entry on the page³. The development of the visual form of

accounting documents is part of the research on the history of accounting.

4. Account books give an insight in everyday life and its economy which can be researched only when the variety of existing entries is transferred into a taxonomy. Medieval and early modern accounts used variety of taxonomies to structure the data. All this can only be accessed by researcher with a thorough encoding of the “content”.

Traditionally this multiple interest in the account books led to a decision on one preferred approach. On paper the text could be printed economically only either as tables or as full transcriptions.

The current state of digital edition theory promises to solve this problem: an edition as a digital resource can include several layers of interpretation and leave the decision on the presentation to the user. But a brief survey done on relevant digital resources⁴ shows these possibilities are rarely exploited in full.

The paper discusses some reasons for this situation, in particular the influence of the dominant schools in digital scholarly editing and the resulting models which focus on textual variance, representation of text and documentary edition. Syd Baumann and Kathryn Tomasek have already suggested changes in the de-facto standard for encoding of historical documents the TEI⁵. They propose an element to describe transactions by referring to the textual parts describing the business facts. An alternative standard for the encoding of business facts can be found in XBRL⁶ which offers a flexible methodology and in the “Global Ledger”-module a taxonomy of basic business facts which can be transferred to many historical documents. While the TEI guidelines are good to encode the accounts on the documentary and linguistic level, XBRL seems to offer a good base to model the economic content. A simple bookkeeping ontology would have to include the following facts: gl-cor:entryDetail(gl-cor:account, gl-cor:signOfAmount(bk:i, bk:d), gl-cor:amount(tei:measure(tei:quantity, tei:unit), gl-cor:debitCreditCode), bk:price(tei:measure[x], tei:measure[y]), bk:transaction(bk:when, bk:where, bk:who, what, gl-cor:entryDetail[bk:debit], gl-cor:entryDetail[bk:credit]) which represents the single entry (gl-cor:entryDetail), their organization in bookkeeping (bk:account, gl-cor:debitCreditCode), the central economical informations (amounts, increase/decrease, prices) and the social act of transaction (bk:transaction).

Departing from the theoretical considerations presented by Manfred Thaller at the DH2012⁷ the paper tries to develop a RDF-model which integrates the multiple layers how an account book can be conceptualised in a digital scholarly edition (text as image, as trace, as language and as meaning). The model is based on references between an URI for the physical object, the images of the object, the transcription and the bookkeeping facts (see fig. 1).

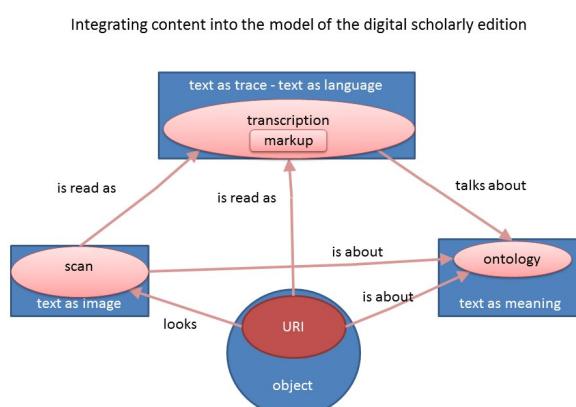


Fig. 1: conceptual model for stand-off markup of digital scholarly editions

The model can be aligned with upper level ontologies like the CIDOC-CRM⁸ as the basic entities relate to linguistic objects

(crm:E33, text as language), to images (crm:E38, text as image), to inscriptions/marks (crm:34/crm:37, text as trace) and to facts which can be described as events (crm:E5, business transactions as the meaning of the accounts), although substantial intermediate steps have to be included (e.g. price relations, measurements and monetary values as subclasses of propositional objects, crm:89). It allows like XBRL the inclusion of taxonomies for commodities or for types of transaction.

The model can be serialized with the help of the feature structures in TEI and converted into explicit RDF with simple XSL which includes the transformation of DOM relations (e.g. ‘contains’) into ontology statements (e.g. <list ana="bk:d"><item ana="bk:entry">For <seg ana="bk:account">wood</seg> <measure ana="bk:amount" quantity="10" unit="lb">x lb</measure></item></list>). First experiences in a project to create a digital edition of a whole series of early modern city accounts show that the model can be used efficiently when the encoding of this kind of structures is supported with a TEI customisation which helps the transcribers to replace repetitive and verbose code with simple XML tags (like <r:lb> for <measure type="currency" unit="lb"> or <r:e> for <item ana="bk:entry">) and when repetitive tasks like the transformation of roman numbers can be included in the XSL transformation from customised code to TEI.

The paper will demonstrate the application of the model to several late medieval account books in a TEI serialization⁹. The integration in the GAMS repository infrastructures¹⁰ allows showing possible functionalities of digital editions based on the model. This includes joint operations on the XML text stored in a Fedora Commons repository and the RDF representation stored in a triple store.

References

1. Die Rechnungsbücher der Stadt Luxemburg, bearb. v. Claudine Moulin u. Michel Pauly, z.Zt. 6 Hefte, Luxemburg 2007 - 2012 (Schriftenreihe des Stadtarchivs Luxemburg ...)
2. Mark Mersiowsky: *Die Anfänge territorialer Rechnungslegung im deutschen Nordwesten*. Spätmittelalterliche 2. Rechnungen, Verwaltungspraxis, Hof und Territorium (zugl. Diss. phil. Münster 1992), Sigmaringen 2000 (Residenzenforschung 9); Georg Vogeler: Spätmittelalterliche Steuerbücher deutscher Territorien, Teil 1: Überlieferung und hilfswissenschaftliche Analyse, in: AfD 49 (2003), S. 165-295, Teil 2: Funktionale Analyse und Typologie, in: AfD 50 (2004), S. 57-204.
3. Federigo Melis (1950): *Storia della Ragioneria. Contributo alla conoscenza e interpretazione delle fonti più significative della storia economica*, Bologna; Franz-Josef Arlinghaus: Bookkeeping, Double-Entry Bookkeeping, in: Medieval Italy. An Encyclopedia, hg. v. Christopher Kleinhennz, New York 2004, S. 147-150; Basil S. Yamey: Scientific Bookkeeping and the Rise of Capitalism, in: EHR N.S. 1 (1949), S. 99-113.
4. Die Cameralia des Stadtarchivs Regensburg, Bd. 3: Ausgaben der Stadt Regensburg 1393 - 1394, ed. by Heidrun Boshof (Fontes Civitates Ratisponensis) www.fcr-online.com/editions/c03/index.htm ; Die mittelalterlichen Schuld- und Rechnungsbücher des Deutschen Ordens um 1400. Eine synoptische Edition im Internet, ed. by Christina Link u. Jürgen Sarnowsky, Hamburg 2008, www.schuredo.uni-hamburg.de; Les comptes des consuls de Montferrand (1273-1319)2006 (Éditions en ligne de l’École des Chartes, volume16), éd. R. Anthony Lodge 2006 elec.enc.sorbonne.fr/montferrand/ ; Comptes de châtellenies, http://www.castellanies.net; Comédie-Française Register Project, http://web.mit.edu/hyperstudio/cfr; Open domesday. The first free online copy of Domesday Book, ed. by Anna Powell-Smith, J.J.N. Palmer, Univ. of Hull http://www.domesdaymap.co.uk; The Alcalá account book project, National University of Ireland, Maynooth 2008, http://archives.forasfeasa.ie; Henry III fine rolls Project, King’s College London et al., 2007-2011, http://www.finerollshenry3.org.uk
5. Encoding Financial Records for Historical Research, paper presented ad the TEI-MM 2012 in College Station, http://idhmc.tamu.edu/teiconference/program/papers#encfin; Syd Bauman, Transactionography Customized Documentation,

Encoding Historical Financial Records Open Access Library, accessed October 28, 2013, <http://omeka.encodinghfrs.org/items/show/5>; hitepaper 2012, NEH Ref: HD-51224-11, omeka.encodinghfrs.org/items/show/4

6. eXtensible Business Reporting Language, www.xbrl.org, Global Ledger Taxonomy: www.xbrl.org/gltaxonomy

7. **Manfred Thaller**: *What is a text within the Digital Humanities, or some of them, at least?* digital humanities 2012 (Hamburg) www.dh2012.uni-hamburg.de/conference/programme/abstracts/beyond-embedded-markup

8. *Definition of the CIDOC Conceptual Reference Model*, ed. by Nick Crofts, Martin Doerr, Tony Gill, Stephen Stead, Matthew Stiff, December 2011. www.cidoc-crm.org

9. *Rechnungen des Mittelalters und der frühen Neuzeit*, Prototypen, ed. by Georg Vogeler, 2012ff. gams.uni-graz.at/rem

10. *Geisteswissenschaftliches Asset Management System [(Humanities Asset Management System)]*, curated by the Zentrum für Informationsmodellierung – Austrian Centre for Digital Humanities gams.uni-graz.at/context:gams/sdef:Context/get?mode=about

Archaeology in social media: users, content and communication on Facebook

Vosyliute, Ingrida

ingrida.vosyliute@kf.vu.lt

Vilnius University Faculty of Communication

The use of social media tools within digital archaeology helps create an engaging setting for archaeological content, which integrates archaeology into a broader social context of use by connecting scholars, archaeological heritage professionals, and the wider public. Social media offers various opportunities: researchers may find it useful for discovering, using and sharing information, organizations may use it in promoting institutional agendas and communicating with wider audiences, lay people may see it as a platform for participation, and more. More broadly, social media is transforming ways in which we perceive information, with its use becoming an increasingly important skill for researchers. This paper attempts to address issues of social media usage in digital archaeology through the case study of studying Lithuanian archaeology practices on Facebook.

The study of social media use in archaeology is still new as a topic of research, because social media haven't been around long enough to develop clear patterns of use. However, particular research questions, as well as answers, are emerging (e.g., Morris, 2011; Whitcher Kansa & Deblauwe, 2011; Pett, 2012; Richardson, 2012; Sanchez, 2013); nevertheless, discussions of social media in archaeology are still more often discussed in conferences and seminars, and also on blogs, forums, etc. Authors typically acknowledge the importance of social media, and point to successful examples providing evidence that it stimulates communication between researchers, helps information sharing and reaching wider audiences, as well as fosters community engagement and social participation. However, the diversity of existing practices opens new research questions, transcend disciplinary boundaries and challenges established authority structures.

The research project presented here is a case study of archaeological communication on Facebook (currently the most popular site for digital social networking in Lithuania), based on analysis of empirical data from thirty Lithuanian Facebook groups and pages related to archaeology. The study depends on a mixed methods approach, combining digital ethnography, content analysis and social network analysis aspects. Initial analysis revealed that overall activity relies on engaged communities rather than on research institutions, or custodian archaeological organizations, considered to be directly responsible for the creation and curation of digital archaeological content. The scope of the research covers, therefore, a wider landscape of observable social media

practices, by actors including not only research organizations or professional networking groups, but also semi-formal or informal groups. Its objective is to map and understand existing trends, and to provide further insights about new phenomena that emerge from these kinds of interactions.

The paper investigates Facebook profiles of individual users (archaeologists, amateurs) and organizations, specific activities they engage in such as posting, commenting, liking, sharing, etc., and the content that is shared within the network. It seeks to address questions arising from this case study, as well as develop insights for broader research issues, such as:

- Who is using social media in archaeology, and for what reason and purpose? What are the qualitative traits, and in depth profiles, of the most effective users? What is the nature of the shift towards public archaeology and community engagement practices? What is the role of individual archaeologists in social media? Could Facebook contribute to research proper, or be used for academic purposes?

- Do social media shape and change the content itself? How do people use and make sense of these resources? What are the most common kinds of archaeological objects that people share, like and comment on Facebook., and why? How is content influenced by the complex relation between archaeological heritage and society? What is the balance between expert knowledge and amateur perspectives?

- More generally, are we fully aware of the opportunities and challenges brought by social media? What additional value does communication among individuals and institutional structures create? How does this kind of synergy improve knowledge transfer? Does it empower organizations, or people? How is communication carried out? What is the structure of interactions between users? In what way is Facebook-based activity shaped to satisfy the needs of its users?

This paper will, firstly, provide an overview of Facebook use in archaeology by focusing on three core dimensions: users, content and communication. It will then present a detailed composition of Facebook users in Lithuanian archaeology, in an attempt to understand the position of archaeological institutions and archaeologists in social media, as well as reasons for the lack of participation as the case study suggests. Furthermore, it will describe the main types and subject-matter of current digital archaeological content, and discuss how user responses and interactions could influence the way in which we conceptualise and interpret the past. Finally, the paper will present and compare different cases of archaeological Facebook use, and will examine in what manner archaeological heritage operates in digital social media, how it serves institutional and individual needs, and what criteria could enable successful communication.

References

- Morris, J.** (2011). *Archaeologist connecting through social media*. The SAA archaeological record, Vol. 11, No. 1 (January).
- Whitcher Kansa, S. & Deblauwe, Francis** (2011). *User-generated content in zooarchaeology: exploring the "Middle Space" of scholarly communication*. In E. Kansa et al. (Ed.) Archaeology 2.0: New Tools For Communication and Collaboration. USA: Cotsen Institute of Archaeology.
- Pett, D.** (2012) *Use of Social Media Within the British Museum and the Museum Sector*. In Ch. Bonacchi (Ed.) Archaeology and Digital Communication: Towards Strategies of Public Engagement. London: Archetype Publications.
- Richardson, L.** (2012) *Twitter and archaeology: an archaeological network in 140 characters or less*. In Ch. Bonacchi (Ed.) Archaeology and Digital Communication: Towards Strategies of Public Engagement. London: Archetype Publications.
- Almansa Sanchez, J.** (2013) *To be or not to be? Public archaeology as a tool of public opinion and the dilemma of intellectuality*. Archaeological Dialogues, Vol. 20, No. 1 (June).

Digital Humanists Are Motivated Annotators

Walkowski, Niels-Oliver

walkowski@bbaw.de

Berlin-Brandenburgische Akademie der Wissenschaften

Barker, Elton T. E.

etebarker@gmail.com

Open University

Background

One core practice that has traditionally supported humanist research process has been the practice to annotate the object of interest in some way. The handwritten additions to books in dusty shelves of libraries are documenting this very well. Although in digital humanities the concept of object is more problematic, annotations are still a fundamental and even increasingly important principle. For example annotations are used for data enrichment and harmonisation of massive digital collections, to model overlapping semantics and structures in presentation systems, and as the key means of connecting online material in Linked Open Data projects. Annotations are one of the core scholarly primitives of humanities research independent of discipline or media.

The more tools with annotation capabilities that are developed and the more annotations that are published, the more interoperability will be critical. The W3C Open Annotation Community Group and the definition of the Open Annotation (OA) Ontology are important steps to ensure interoperability among digital annotations. Nevertheless some important aspects remain to be done. The OA model describes a comprehensive way of rendering the structure of digital annotations; but its goal is not to reflect the methodological or hermeneutical context of digital annotations beyond what is needed to make the model generic. Moreover, the model claims that every situation where a web-resource can be linked to another web-resource can be modelled as an annotation.

Problem

The more annotations are published in a consistent way as web resources - a process that is promoted by the Semantic Web framework of the OA model - the more the question of interoperability shifts from the representation of structure to context. The problem here is not the use of a predicate to refer to the annotation target rather than to the annotation body; that can be easily resolved by the inclusion of any concept which is dereferenceable between oa:hasBody and oa:hasTarget. The problem is more that annotations which refer to the same object and which use the same concept may have meanings depending on the research context different from those for which the annotation were originally created. Consider the following examples:

- For humanities researchers an algorithm represents an object of interest as much as any other cultural object. That is to say, annotations created in a context where the algorithm is investigated should not be used as expressions about the objects on which this algorithm was tested. As a result, modifications that are made during the investigation ought to reveal something about the algorithm not about the object.
- Annotations about the authorship of Shakespeare that are made in a stylometric analysis proving which works should be attributed to Shakespeare ought to be processed and interpreted differently from authorship annotations that create a catalogue of texts in a project.
- An annotation created in a crowdsourcing context without the use of a formal ontology is suitable for other research questions than the same annotation created by a disciplinary expert who applies a related ontology.

These cases demonstrate that the efficient reuse of published annotations needs a formalized representation of the research

purpose for which the annotation was created in the first place. On the other hand such metadata would also provide a valuable information source in the digital environment for traditional humanities research questions. As Claudine Mouline has put it, contextual data about annotations ought to be an invaluable resource for investigating the social or cultural structures out of which these annotations were made. In one of her examples she gets insights about missionary work in Anglo-Saxonia as well as insights about the social structure of monasteries not only from the content of annotations but also from the context. Annotations made by high-priests can be compared with annotations written by monks if these can be distinguished at the first place. Moulin describes how this knowledge is derived from knowing in which monastery a book was annotated and which tool was used for the annotation because some tools were only used by high-priests. Different purposes for annotations were revealed giving the annotation a specific meaning.

It is true that the first draft of the Open Annotation Ontology attempts to address this issue by introducing the predicate 'oa:motivatedBy', in order to describe the motivation or intention associated with the annotation apart from its content. Nevertheless, the exact use of this predicate is still an open question, especially its relation to the type of annotation (Highlighting, Sticky Note, etc.). Indeed, the community group emphasizes the importance of a general approach for using this predicate and encourages the definition of vocabularies in communities so that proliferation of terms can be avoided. Yet, it is still unsure if the oa:motivatedBy predicate will be kept in the final version of the Open Annotation model as was expressed by Timothy Cole at the DH2013 in Nebraska.

Approaches

The value of maintaining the concept of an oa:motivatedBy predicate may be seen by considering large European infrastructure projects like Dariah and EUDAT which are developing annotation tools as generic components of their infrastructure. The generic structures and models adopted by these projects strongly demand that digital annotation practices and purposes ought to be systematized. At the same time the massive scale of these projects, resulting from their integrated services, offers a variety of possibilities of defining a vocabulary that can be used for oa:motivatedBy in the Digital Humanities. Specifically we suggest three sources:

- The results of the Dariah-DE experts workshop on interoperable annotations for the arts and humanities
- The efforts around the definition of the Scholarly Methods Ontology in Dariah-EU
- The Scholarly Domain Model presented by Stefan Gradmann at the DH2013

Each of these sources represents a different approach to how the motivation of an ontology can be formalized. The methodological part of the June 2013 "Dariah-DE Experts Workshop on interoperable annotations" built a classification upon the evaluation of practices in which digital annotations are used (to collaborate, to review, to enrich, etc.): a DH motivation vocabulary ought to be grounded in general practice and from the annotation point of view. The Scholarly Methods Ontology tries to classify digital research activities according to a methodological framework. Therefore it represents the best possibility of creating a more abstract vocabulary grounded upon theoretical reflections about how the field of DH is organized. The Scholarly Domain Model differs from the former approach because it does not focus on any research activity but on the areas in which humanities research activities take place: these areas are Input, Output, Metadata and Social Context and Research.

To put it more simply: the first classification model relates to the annotation object itself, the second tries to reflect fields of research, while the third introduces an abstract view on knowledge production itself. It would be misleading to select any one of these perspectives as the annotation model, since all of them have advantages and disadvantages. For example, the Scholarly Domain Model is currently too general in terms

of semantics to help overcome interoperability issues for annotations. Besides, these models do not pertain to mutually exclusive ideas: enrichment by annotation in the first model is also suitable to describe the metadata class of the Scholarly Domain Model, while reviewing is located along the input/output axis. Thus, rather than prioritizing one perspective over another, we suggest an ontology concept for oa:motivatedBy that integrates these three models through their overlaps and interconnections, which in turn respects the idea that any ontology should relate annotation patterns to specific research fields.

Perspectives

The problems and approaches presented in the context of annotations and the Open Annotation Ontology address more complex ongoing discussions, such as the need to have a sustainable and expressive concept of provenance for digital research objects. This task is especially difficult in the case of humanities since the frequently non-processual and non-linear elements of humanities research do not sit well with the purely event- and agent-based provenance models created in the e-Science realm. The case of annotation is a good example. The meaning of a result for a humanities-related research question is not made totally transparent by the process or the broader lineage in which it 'happened'. The Open Annotation Ontology suggests the use of the Provenance Ontology which is modeled along these concepts therefore not offering additional help. Our contribution, then, aims not only to show the importance of the oa:motivatedBy predicate and make the first steps towards developing a concept for a related ontology; we also hope to encourage the development of a discriminating idea of provenance in the humanities that is rarely developed at present time.

References

- Simon, R., & Jung, J.** (2011). *The YUMA Media Annotation Framework*. Research and Advanced Technology for, 434–437.
- Sanderson, R., Albritton, B., Schwemmer, R., & Van de Sompel, H.** (2011). *SharedCanvas: A Collaborative Model for Medieval Manuscript Layout Dissemination*. arXiv.org, 1104(2925). Retrieved from arxiv.org/abs/1104.2925
- Simon, R., Barker, E.T.E. and Isaksen, L.** (2012): *Exploring Pelagios: a visual browser for geo-tagged datasets*. E. Agirre, K. Fernie, A. Otegi and M. Stevenson (eds.), International Workshop on Supporting Users' Exploration of Digital Libraries. Cyprus, 29-34.
- Unsworth, J.** (2000). *Scholarly Primitives: what methods do humanities researchers have in common, and how might our tools reflect this*. In Humanities Computing: formal methods, experimental practice symposium, King's College, London. www.w3.org/community/openannotation/
- Sanderson, R., & Van de Sompel, H.** (2013). *Designing the W3C Open Annotation Data Model*.
- Moulin, C.** (2010). *Vom mittelalterlichen Griffel zum Computer-Tagging. Zur sprach- und kulturgeschichtlichen Bedeutung der Annotation*. In Akademien der Wissenschaften und der Literatur Mainz. Jahrbuch 2012 (pp. 84–99). Mainz. www.w3.org/community/openannotation/wiki/Open_Issues de.slideshare.net/gradmans/20130719-dh2013-beyondinfrastructure www.w3.org/TR/prov-o

The dog that didn't bark: A longitudinal study of reading behaviour in physical and digital environments

Warwick, Claire

c.warwick@ucl.ac.uk
UCL

Mahony, Simon

s.mahony@ucl.ac.uk
UCL

Rayner, Samantha

s.rayner@ucl.ac.uk
UCL

Team, The INKE

www.inke.ca
INKE

1. Introduction

The following proposal presents the results of a longitudinal study of reading in both digital and physical environments carried out as part of the INKE project, by researchers at the UCL Centre for Digital Humanities from 2009-2013. Our aim was to understand how different digital devices and physical, printed publications are used, and integrated into reader behaviour. It took place at a time of rapid change in the technology of reading in digital environments: the Kindle, iPad, and other tablet computers were launched, and quickly became popular, during the period of the study. We might hypothesise, therefore, that such a period would usher in the transformation of the reading experience, and wholesale move from paper to digital devices and ebooks which had been predicted in the literature.^{1 2} Our study of reading behaviour provides evidence against which to test such claims.

1.1. Research context

Reading has been studied by cognitive scientists (see for example the journal *Scientific Studies of Reading*); those interested in literary reading³; and in how children learn to read (for example the very widely cited study by Wagner et al. 1997) but the way in which we integrate reading into our lives and behaviour is relatively little understood. Two different broad types of reading may be observed: reading to gather information, and immersive reading. The first, described by Vandendorpe as ergative reading, is used to find information in printed texts or digital documents. This is, he argues, well suited to digital environments, which make information seeking faster and more efficient.⁴

Immersive reading is a complex interaction with a text, which we may read for pleasure, or which literary scholars may study and read closely such as novels or poetry. This is more difficult to study, entailing understanding of complex language, extended narrative, rhetorical devices etc.⁵ Yet, contrary to Vandendorpe's predictions, it remains as common as ergative reading: millions of people engage in such an activity every day, and many of them use digital devices, such as tablets. Yet, arguably, such technologies do little more than mimic the affordances of the printed book, and some users still find it hard to read long documents in any form other than paper. Thus it is still unclear whether digital environments offer significant advantages to readers of long, complex narratives, as compared to print.

Our study was therefore aimed to understand how users carry out both kinds of reading; what kind of device they use for which tasks; what they enjoy and what frustrates them about such devices; and what features and affordances they would like to see in new digital environments. The results of our study provide an enhanced understanding of reading behaviour, at a time of remarkable technological change, and aims to evaluate the effects of such rapid developments on readers and their views about reading.

2. Methodology

This study was intended to capture change over a lengthy period of time and is, as far as we are aware, unprecedented, in terms of a study of adult readers, although there are numerous longitudinal studies of reader development in children: recent examples include studies by Nation and Hulme,⁶ and Oakhill and Cain.⁷ Its longitudinal nature is important: while other studies have been carried out into the practice of digital reading (for example, Gerlach and Buxmann's study)⁸, they provide snapshot of a user population. We argue that only by repeating the same task over several years is it possible to determine how, and whether, reader behaviour changes in response to different technologies. It was a cross-sectional study: the research activity remained the same over the course of the project, but we did not track one user group over the study period.⁹

We chose to study a group of Masters students at the UCL Department of Information Studies, because they were adult, needed to read widely as part of their course, and were, typically, from a background in the humanities, thus we hoped they would also be likely to read for pleasure. We also integrated the study into the teaching and learning of the Electronic Publishing module, as a way of introducing the students to the practice of action research, and enquiry-based learning early in their course. Thus the study benefited not only our research, but also the student learning experience. Each group also included a significant number of non-UK students, many of them non-native speakers of English: this therefore allowed us to determine whether nationality or native language had any effect on reading behaviour.

Each participant was asked to keep a diary of their reading for a week. We asked them to note the time period in which they read, where they were, what technology they were using (making it clear that printed books, newspapers and magazines are reading technologies), and to note any problems they encountered or other reflections that they might have. Diary studies provide information about behaviour patterns, and attendant problems when using different technologies, and we have previously used them to study of field archaeologists' use of digital technologies¹⁰. We then held a group discussion session, in which the students reported back on their experiences, and responded, firstly in small groups, then in plenary discussion, to a series of questions about reading in different physical and digital environments. This allowed us to capture their views, and suggest a series of requirements and affordances for future digital reading devices. The reading diaries were then collected and content analysis techniques used to identify patterns in the data. As the study progressed, themes were compared with data from previous years to identify change over time

3. Findings

We found that our users had a very complex, almost instinctive, understanding of the affordances of different reading devices and environments. They had a wide repertoire of contexts for reading; physical setting such as the bus to work, the office, the sofa or, of course, in bed; or devices, such as phones, cereal packets, free newspapers, computer monitors, printed books, and, with growing frequency as the study progressed, tablets and e-readers. They typically read for relatively short periods, and moved frequently from one device or context to another, choosing print or digital media depending on their task; physical setting; whether they were on their own, or with colleagues; and the technologies available; rather than having dogmatic preferences for one type of device or environment over another.

Despite the rapid development of tablet and e-reader technologies, we were surprised to find that their views about design and requirements changed relatively little. While appreciating the potential of tablets for information seeking, and for storage and fast retrieval of a large number of books or documents, users still felt an emotional attachment to the book as an object. There was also no significant difference between the behaviours and views expressed by students of different nationalities. Users were well aware of the affective

aspects of books, such as the feeling of paper under the hands, and its flexibility and warmth, as opposed to the cold, uninviting feeling of plastic or metal, and valued attractive book design, and the distinctive smell of newsprint or rare books. Despite the numerous experiments with annotation, navigation and bookmarking technologies in digital environments, users also expressed a strong preference for annotating physical copies, and valued the ability to flip through pages of a printed book as a way of orientating themselves.

As the study progressed more of our users seemed comfortable with reading long documents, including entire novels, in digital form, but it appears that those who do so are heavy users of tablets or e-readers, perhaps as a result of being early adopters. Surprisingly large numbers of respondents still prefer to read in print, and many of these are relatively light users of digital devices. Many of these users reported that it was difficult for them to feel the sense of flow- that is being absorbed in the narrative, and unaware of the device on which it is being read- when using digital devices as opposed to reading in print. It is difficult, at this stage, to tell whether the ability to read at length on a tablet is something acquired through practice, and enthusiasm for this medium, or whether users who are unenthusiastic about digital reading do not persist, and thus never acquire the habit of use and resulting enjoyment of flow.

4. Conclusion

The particular significance of our study results not only from individual findings, some of which others have also noted,¹¹ but in the fact that, despite our initial hypothesis, we found a remarkably stable set of behaviours and user requirements irrespective of rapid technological change. Users understand the advantages of digital reading devices, but still value the tactile, visual and affective aspects of the printed book. It appears that the printed book is not as easily modelled in, or replaced by, digital environments, as had once been thought: doing so remains a significant challenge for future scholarship.

References

1. Nunberg, G. (ed.) (1996) *The Future of the Book*. Berkley, University of California Press.
2. Birkerts, S. (1995) *The Gutenberg Elegies: The fate of reading in an electronic age*. London, Faber.
3. Miall, D.S., Kuiken, D. (2002). *A feeling for fiction: Becoming what we behold*. Poetics, 30, 221-241.
4. Vandendorpe, C. (2011) *Some Considerations About the Future of Reading*. Digital Studies/Le Champ Numerique, 2, 2 (www.digitalstudies.org/ojs/index.php/digital_studies/article/view/186/250)
5. Warwick, C. (2004) *Print Scholarship and Digital Resources*. In Schreibman, S., Siemens, R., Unsworth, J. (Eds.) *A Companion to Digital Humanities* (pp. 366- 382). Oxford, Blackwell.
6. Nation, K. and Hulme, C. (2011) *Learning to read changes children's phonological skills: evidence from a latent variable longitudinal study of reading and nonword repetition*. Developmental Science, 14, 4: 649-659
7. Oakhill, J., and Cain, C. (2012) *The Precursors of Reading Ability in Young Readers: Evidence From a Four-Year Longitudinal Study*. Scientific Studies of Reading, 16, 2.
8. Gerlach, J., Buxmann, P. (2011) *Investigating the acceptance of electronic books- The impact of haptic dissonance on innovation adoption*. ECIS 2011 Proceedings. Paper 141. aisel.aisnet.org/ecis2011/141
9. Ruspini, E. (2002) *An Introduction to Longitudinal Research methods*. London, Routledge. Chapter 1.
10. Warwick, C., Terras, M., Fisher, C., Baker, M., O'Riordan, E., Grove, M., Fulford, M., Clarke, A., Rains, M. (2009) *iTrench: A study of user reactions to the use of information technology in field archaeology*. Literary and Linguistic Computing 24, 2: 211-224.
11. Jabr, F. (2013) *The Reading Brain in the Digital Age: The Science of Paper versus Screens*. Scientific American, April 11, 2013 www.scientificamerican.com/article.cfm?id=reading

paper-screens&&goback=.gde_104765_member_233183076#%21

Ideas, Events and Actions: The Digital Humanity Study of the Concept Formation in Modern China

Wen-huei, Cheng

National Chengchi University (Taiwan), TW

Jui-sung, Yang

National Chengchi University (Taiwan), TW

Wei-Yun, Chiu

National Chengchi University (Taiwan), TW

Chao-lin, Liu

National Chengchi University (Taiwan), TW

Guan-tao, Jin

National Chengchi University (Taiwan), TW

Qing-feng , Liu

National Chengchi University (Taiwan), TW

1. Introduction

How to figure out the formation process of concepts has long been a significant yet elusive problem in Humanities studies. In order to better understand this problem, we have drawn on insights from History of Ideas (Lovejoy, 1936; Pocock, 1898; Skinner, 2002; Jin et al., 2008), Conceptual History (Koselleck, 2002). Key word studies (Williams, 1983) and computational linguistics (Wittgenstein, 1953; Austin, 1962; Deignan, 2005). More importantly, we have also employed data analytics (Liu et al., 2011), Zipf's law, and statistical methods (Jin et al., 2012). As a result, in terms of methodology, we have come up with a new approach of Chinese historical studies, which explores temporal analysis of keywords and their collocations and promises to outline the trajectory of conceptual formation more accurately. By means of this new approach, we have already investigated closely the formation processes of three concepts ("ism", "Chinese", "Chinese labor") in modern China with good results (Chan et al., 2011; Jin et al., 2012). These studies demonstrate the potential power of DH methods for the study of ideas. Encouraged by the fruits of previous studies, we now try to look into the formation process of three important keywords in the construction of modern Chinese national identity, i.e., "guojia (nation-state)", "sovereignty", and "tongbau (siblings)." By utilizing DH methods, we aim to analyze the processes, structures and patterns of formation of concepts in critical phases, and to portray the dynamic schema of the interactions between ideas, events and actions. In the future, we will apply more computational linguistics methods in our study of concept formation process, in order to understand more clearly the birth and evolution of key ideas.

2. Body Paragraphs

Given the fact that during the modernization process of China, the formation of concepts concerning national identity such as "nation-state", "sovereignty", and "siblings" is a very important cultural issue; we need to have new approaches and perspectives to study it, especially in the face of the new age of big digitalized corpora. By handling the huge historical corpora with DH methods, we are able to better grasp the interactive processes and patterns between concept formation, important social events and actions of different groups in modern China.

To investigate the concept formation of "nation-state" in modern China, we choose "Xinmin Congbao" and "Xin Qingnian" as two major sources for analysis. The former was published in the final years of late imperial China, while the later

in the early Republican period. By identifying frequent keywords with the PAT-Tree method and computing statistics of these keywords and their collocations, we analyzed the differences between the two sources in terms of their interpretations of the "nation-state" concept. In addition, we explored the complicated relationship between the changing definitions of "nation-state" and important social events and actions. As a result, we have found out that, before 1911 while the formation of "nation-state" concept was inchoate, this concept was mainly embedded in the concepts of citizen and individual, and became popular mainly as a result of civic education. However, after 1911 when "Xin Qingnian" was published, the "nation-state" concept has become very widespread. In particular, along with the breakout of the World War I, under the "party-state" system in modern China, the "nation-state" concept became highly ideologized. Textually speaking, it frequently appeared with terms such as "class" and "capital" in various discourses, relating closely with class revolution and economic revolution as well.

On the other hand, by mapping out the linguistic development of "sovereignty" in modern China, we have outlined three stages of the concept formation of this term. During the first stage (1864-1898), the Western definition of "sovereignty" was selectively interpreted in late imperial China. While "sovereignty" was introduced as a modern idea representing the right of a nation, the Chinese emperor was still conceived as the only master of the sovereignty. In other words, the modern (Western) sense of sovereignty was partially accepted for its instrumental function for dealing with international affairs. During the second stage (1899-1915), China was trying to exercise sovereignty as a modern state. China has then transformed from a dynastic empire into a nation-state and more competitions and negotiations between China and other nation-states took place. Thus, the modern idea of sovereignty became much more popular. During the third stage (1916-1924), the sovereignty concept underwent a drastic transformation. The "party-state" system became dominant and produced a new set of moral-political ideologies, which came to define the sovereignty of a nation-state under the party-state framework and claimed it should be under control by the party-state system.

As well known, the construction of modern Chinese identity is closely related to the idea of "tongbau," which literally means that every member of the nation is blood kin, united by blood bond. We have utilized DH approaches to explore how and why this "familial" term has not only transformed people's loyalty from families into the nation but also "naturalized" the sense of solidarity and patriotic love. Our initial investigation of the origin of the modern meaning of "tongbau" has revealed that modern Japan might be the source. We have found that the highest frequency of using the term "tongbau" in its modern meaning appeared in three famous journals (Jiangsu, Xinmin congbao, Qingyibao) published after 1898 in Japan by the exile Chinese intellectuals and students there. This finding is very significant. First of all, the timing itself deserves to be explored further. As well known, the failure of the 1898 political reform in China stimulated many intellectuals to seek popular support for political reform instead. In other words, the rise of "tongbau" discourse in modern China was closely related to the political development in late Qing. Secondly, the fact that "tongbau" discourse originally appeared in Chinese journals published in Japan also provides a piece of solid and interesting evidence to further testify the complicated relationships between modern China and Japan, especially regarding the construction of modern national identity.

3. Conclusion

The birth of a key concept is related to important events and actions, and the formation of a new concept will inevitably bring out changes of values. We have utilized new DH approaches, computational linguistics, analysis of the formations of concept terms, in order to achieve the following goals: (1) outlining the formation structure of concepts, (2) probing how China transformed from a traditional dynastic empire into a modern nation-state, (3) examining the contour of the changing identity from a "subject" into a "citizen" among the Chinese people, and

(4) exploring how key Western concepts were translated and reinterpreted in China in the past 150 years.

The target sources for our study are mainly important journals published during the late Qing and early Republican period: "Qingyibao" was published in Japan after 1898 by a pro-emperor group. It was later succeeded by "Xinmin Congbao." The editor-in-chief and mastermind for both journals was the important thinker Liang Qichao. Hence, these two journals vividly manifested the changing attitudes and ideas of the pro-emperor group. "Xin Qingshian" was an important journal during the early Republican period, indicating the intellectual trend of the time. With the help of DH methods, we were able to study them in a new manner and therefore reached our research goals aforementioned.

It should be mentioned that, since modern Chinese newspapers served as the platforms for enlightenment and reform, they featured both the elite and the mass. Although it is impossible now for our studies to include all variables, we will in the future try our best to use "media," "readers," "editors," "political stereotypes" or "newspapers position" as our major variables for further investigation. Moreover, in order to strengthen our digital skills, we will use statistical keyword extracting analysis and co-occurrence word cluster statistical method, as well as social network analysis, citation analysis and spatial-temporal analysis. After all, the main characteristics of our studies lie in using DH methods to do Chinese text analysis. We are among a small pioneer group attempting to experiment DH approaches in Chinese studies. We are convinced that the results of our studies are very meaningful and they will provide rich resources of reference for applying DH approaches in textual analysis in other languages.

Based on what we have done so far, we will further employ theories from cognitive sciences, allegory theory, statistics methods, thinking about the possibility of new breakthroughs of DH approaches. We are confident that our new methods will shed new light on current DH studies. The unique perspectives of DH studies, which differ from the traditional historical approaches, will offer new methods and open up new problem domains, making important paradigm shift in Humanities studies.

References

- Austin, J. L.** (1962). *How to do things with words*. Cambridge: Harvard University Press.
- Chan, C.-Y. and Wang, N.-X.** (2011). *The "isms" of Digital Humanities*. In Hsiang, J. (ed), New Approaches to Historical Studies. Taipei: National Taiwan University Press. (In Chinese)
- Chen, C.-L. and Liu, C.-L. and Chang, Y.-C. and Tsai H.-P.** (2013). *Opinion mining for relating multiword subjective expressions and annual earnings*. US financial statements, Journal of Information Science and Engineering, 29(4):743–764.
- Deignan, A.** (2005). *Metaphor and corpus linguistics*. Amsterdam, The Netherlands; Philadelphia, Pa. : John Benjamins.
- Huang, W.-J. and Liu, C.-L.** (2013). *NCCU-MIG at NTCIR-10- Using lexical, syntactic, and semantic features for the RITE tasks*. Proceedings of NTCIR-10. Held 18-21 June 2013.
- Jin, G.-T. and Chiu, W.-Y. and Liu, C.-L.** (2012). *Frequency analysis and applications of "co-occurrence" phrases: The origin of the 'Hua-Ren' concept as an example*. In Hsiang, J. (ed), Essential Digital Humanities: Defining Patterns and Paths. Taipei: National Taiwan University Press. (In Chinese)
- Jin, G.-T. and Liang, Y.-Y. and Yu, Y.-S. and Liu, C.-L.** (2012). *Lexical statistics and key terms in Chinese documents of the early 20th century*. Proceedings of the Fourth International Conference of Digital Archives and Digital Humanities. Held 29-30 November, 2012. (in Chinese)
- Jin, G.-T. and Liu, Q.-F.** (2008). *Study of the History of Ideas: Formation of Modern China's Important Political Terms*. Hong Kong: The Chinese University of Hong Kong Research Centre for Contemporary Chinese Culture. (In Chinese)
- Koselleck, R.** (2002). *The Practice of Conceptual History: Timing History: Spacing Concepts*. Todd S. P. et al., (trans). Stanford: Stanford University Press.
- Lakoff, G. and Johnson, M.** (1989). *Metaphors we live by*. Chicago: University of Chicago Press.
- Liu, C.-L. and Jin, G.-T. and Liu, Q.-F. and Chiu, W.-Y. and Yu, Y.-S.** (2011). *Some chances and challenges in applying language technologies to historical studies in Chinese*. International Journal of Computational Linguistics and Chinese Language Processing, 16(1-2):27-46. <http://www.aclclp.org.tw/clclp/v16n12/v16n12a3.pdf>
- Lovejoy, A. O.** (1936). *The Great Chain of Being: A Study of the History of an Idea*. Cambridge: Harvard University Press.
- Pocock, J. G. A.** (1989). *Politics, Language, and Times: Essay on Political Thought and History*. Chicago: University of Chicago Press.
- Skinner, Q.** (2002). *Visions of Politics*, vol. I, Regarding Method. Cambridge: Cambridge University Press.
- Williams, R.** (1983). *Keywords: A Vocabulary of Culture and Society*. New York: Oxford University Press.
- Wittgenstein, L.** (2001[1953]). *Philosophical Investigations*. Anscombe E. (trans). Oxford: Blackwell Publishing.

Lacuna Stories: Building an Annotation Platform for Historical Thinking

Widner, Michael

mikewidner@stanford.edu
Stanford University Libraries

Johnsrud, Brian

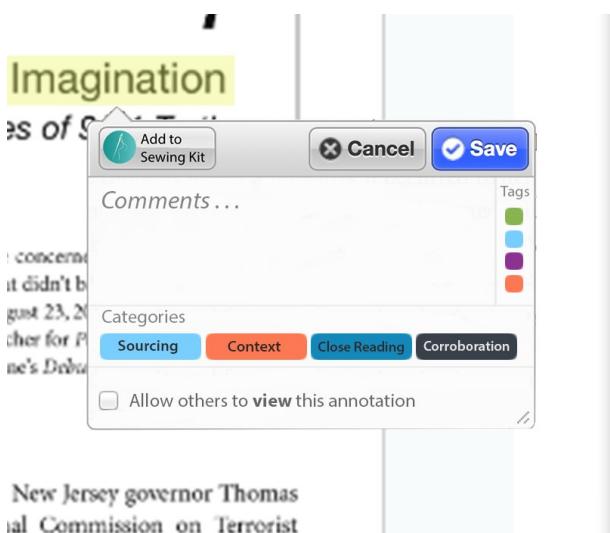
johnsrud@stanford.edu
Stanford University

Participatory and collaborative sense-making of complex phenomena is central to productive learning and knowledge work in today's information-rich world.¹ The Lacuna Stories Project creates an exploratory, interactive, and collaborative online space where users can research and discuss significant historical events like 9/11. Lacuna Stories draws together primary source documents, fiction, scholarship, wikis, and user-generated forums and blogs. The online space extends current digital annotation software with functionality that encourages skills such as historical thinking, close reading, and comparison of media and sources concerning 9/11. When approached independently, individual sources, genres, and media inevitably fall short of stitching together the "whole story." The Lacuna Stories Project's diverse, multimedia environment provides tools for instructors, students, and the general public to "mend" the gaps in our knowledge of major historical events in order to develop their own narratives. User data generated through the site's design also will allow researchers to compare and better understand reading and engagement behaviors of students; along with other forms of user experience research and assessment, this research also provides directions for improving the platform and instructional materials.

The Lacuna Stories Project is a cross-disciplinary collaboration that includes faculty whose research interests span literature, pedagogy, historical thinking, public humanities, and platform studies. This short paper will describe the research and pedagogical goals of the Lacuna Stories Project as well as the technological innovations developed to support these goals. By the time of the conference, Amir Eshel (PI) and Brian Johnsrud (Project Manager) will have taught a course during Stanford's Winter Quarter in which students use the Lacuna Stories Platform in and outside the classroom. Johnsrud and Michael Widner (Technology Director) will also have taught a course in the same quarter on building in the digital humanities that will use Lacuna Stories as its primary example. This paper will touch upon how the prototype encouraged student learning and collaboration by presenting our data gathered from student use, interviews, and focus groups. By the time

of the conference will also have piloted Lacuna Stories for a test-group of up to 30 public users without a college degree to compare experiences of different kinds of users in and out of the classroom, with and without formal academic training in approaching different kinds of historical texts and media.

Compared to other annotation and archive projects, the Lacuna Stories platform provides three key innovations. First, it creates an integrated multimedia environment that encourages the development of core skills for learning and knowledge work: navigation, critical reflection, linking, synthesis, and collaborative sense-making. Second, no existing digital annotation tool connects multiple types of media together to compare and generate new narratives. We are coordinating with MIT's HyperStudio to add this functionality to Annotation Studio (www.annotationstudio.org) and incorporate it into an ecosystem of digital tools for collaborative learning. Third, Lacuna Stories will provide a novel, curated set of diverse 9/11 resources for users to engage with and connect in innovative ways.



New Jersey governor Thomas
al Commission on Terrorist

Fig. 1: Annotation Functionality

Lacuna Stories is also a platform that fosters good habits of close reading and thinking historically, whether the users are students, researchers, or members of the general public. Within the humanities specifically, the interactive, multimedia functionality of Lacuna Stories goes beyond simply replacing print reading and viewing practices; rather, it creates new and innovative experiences for engaging with various texts and media that reflects the networked state of knowledge today. The platform thus builds upon the work done by Sam Wineburg, another member of the project team and Margaret Jacks Professor of Education and (by courtesy) of History at Stanford, to promote historical thinking (sheg.stanford.edu). Wineburg's work to date, however, has been focused on print resources in high school classrooms. Lacuna Stories will bring Wineburg's deep engagement with these matters to digital resources and make them available in the university setting.

The project also seeks to develop an inclusive, empowering, and engaging open-source platform to gather and encourage these responses in a generative and reparative mode. The site aims not to develop a fully coherent or conclusive "truth" of the event, but to encourage the cognitive and imaginative work that inspires responses to and stories about the event, its complexity, and its diverse meanings. Lacuna Stories' subtitle, "mend the truth," refers to site's ability to connect in novel ways the different text, media, and user-generated content. One of our primary contributions to the development of Annotation Studio will be to enable all aspects of the available resources—from images to individual words, lines, or documents to user-generated content—to be archived or collected by registered users into their personal "sewing kit," which provides users a workspace for the collection, connection, and annotation of materials relevant to their learning and scholarship.

We will incorporate and extend pre-existing open-source projects for the platform, most notably Annotation Studio from

MIT's Hyperstudio group. Annotation Studio is an exemplary, user-centered tool for digital humanities work, currently under active development in conjunction with a wider set of projects to develop shared standards for annotation of multimedia content, a key requirement for widespread adoption both in and outside of classrooms. We are working closely with the Annotation Studio team to ensure that the innovations developed for Lacuna Stories make their way back into the central code base and are thus available to the broadest possible number of users and institutions. One of the core technologies powering Annotation Studio is the javascript library Annotator.js (okfnlabs.org/annotator), a project of the Open Knowledge Foundation that is quickly becoming one of the most popular annotation technologies. By focusing first on developing extensions to Annotator.js, the Lacuna Stories Project ensures that our work will be useable in Annotation Studio and by any other projects that use Annotator.js.

We are also working to bring the functionality provided by Annotator.js into the Drupal platform that powers much of the rest of the Lacuna Stories site. Once this work is complete, users will be able to annotate not only texts available through Annotation Studio, but also blog posts, wiki entries, and any other content in a single, integrated environment. In our talk, we will discuss some of the challenges in the prototype phase for this work and our reasons for using Annotation Studio as a replacement until the Drupal work is complete.

O'Malley and Rosenzweig argue for the growing importance of the web generally because it allows for communication and exchange of divergent interpretations of the past. The web demonstrates how "meaning emerges in dialogue and that culture has no stable center, but rather proceeds from multiple 'nodes'" (154).² Being able to create links between annotations and sources and annotate the quality of those connections is central to the academic process of synthesizing information across documents and reflects the natural associative mechanisms that are central to deep learning.^{3 4}

This functionality, however, does not exist in any current digital annotation tools. Lacuna Stories seeks to change this fact, with a tool that is scalable for use in a multitude of sense-making settings. Lacuna Stories will allow users to create categories and links between items in their kit, such as connecting a line from a novel with a paragraph from a user-submitted story, a forum discussion thread, and a section from the 9/11 Commission Report. These links can additionally be connected to a larger theme as described by the user; there can also be a shared set of themes developed collectively by the group or by an instructor. Social learning will be enabled through opt-in sharing functionality, where other learners can view and extend links and notes. Such open linking from users' digital "sewing kits" exemplifies the idea that connections among narratives can be made quickly and simply, empowering users to "mend," create, or share meaningful associations. Moreover, this aspect of the project responds to the work done by Fred Turner (Associate Professor of Communication), another faculty member of the team, to create digital humanities projects that are public-facing and that encourage community engagement.

Lacuna Stories is, then, a platform driven by the complementary research interests of a cross-disciplinary team of faculty made possible through technological innovation based on existing, open source tools. Although this paper will focus primarily upon the technology used and plans for future work, a secondary focus will be how the tools and innovations are grounded in research and pedagogy and how these interests influenced our technology choices and strategy.

References

1. U.S. Department of Education. (2010). *National Education Technology Plan*. Retrieved March 8, 2013, from www.ed.gov/technology/netp-2010
2. O'Malley, M. and Rosenzweig, R. (1997). *Brave New World or Blind Alley? American History on the World Wide Web*. Journal of American History 84(1): 132–55.

3. **Tashman, C.S. and Edwards, W.K.** (2011). *Active reading and its discontents: the situations, problems and ideas of readers*. In Proceedings of CHI '11, 2927–2936. CHI '11. ACM.
4. **Marshall, C.** (2005). *Reading and Interactivity in the Digital Library: Creating an Experience that Transcends Paper*. In Digital Library Development: The View from Kanazawa, D. Marcum and G. George, Eds.

Building the social graph of the History of European Integration: A pipeline for the Integration of Human and Machine Computation

Wieneke, Lars

lars.wieneke@cvce.eu

Centre Virtuel de la Connaissance sur l'Europe

Sillaume, Ghislain

Centre Virtuel de la Connaissance sur l'Europe

Düring, Marten

Centre Virtuel de la Connaissance sur l'Europe

Pasini, Chiara

Dipartimento di Elettronica, Informazione e Bioingegneria

Fraternali, Piero

Dipartimento di Elettronica, Informazione e Bioingegneria

Tagliasacchi, Marco

Dipartimento di Elettronica, Informazione e Bioingegneria

Melenhorst, Marc

Delft University of Technology

Novak, Jasminko

European Institute for Participatory Media

Micheel, Isabel

European Institute for Participatory Media

Harloff, Erik

European Institute for Participatory Media

Garcia Moron, Javier

Homeria Open Solutions S.L.

Lallemand, Carine

Public Research Centre Henri Tudor

Vincenzo, Croce

Engineering Ingegneria Informatica S.p.a.

Lazzaro, Marilena

Engineering Ingegneria Informatica S.p.a.

Nucci, Francesco

Engineering Ingegneria Informatica S.p.a.

CUbRIK and the History of Europe App

The integration of human expertise and machine computation enables a new class of applications with significant potential for the digital humanities. So far this potential remains largely untapped due to the severe requirements of such projects: The implementation and integration of advanced algorithms requires specialized know-how and the final users from the humanities are challenged with defining unprecedented tasks for methods which haven't emerged yet. The FP-7-funded research project CUbRIK (www.cubicproject.eu) implements and integrates research in computer science, the design of human-computation tasks, data visualization, social engineering and the humanities.

In the proposed presentation we would like to showcase one of CUbRIK's case studies, the demo of the History of Europe application. The application introduces an effective interface

to access collections of historical sources and to discover links among and entities within them. Upon completion CUbRIK will offer an innovative approach to human-enhanced time-aware multimedia search by synthesizing research in computer science, crowdsourcing and gamification. We will conclude the presentation with an outlook on the future development of the application.

Humanist-machine interaction

The History of Europe (HoE) application is based on a curated collection of more than 3000 images, representing the main events and actors in the history of the European integration. The collection is curated and hosted by the Centre Virtuel de la Connaissance sur l'Europe (CVCE). In a first step, an image indexation pipeline identifies the location of individual faces in the photographs. The location of these faces is verified by a crowd of "click-workers" with no specific training who evaluate for each recognized face if the depicted image shows a human face or not. Following the face verification process, an automatic face recognition process is triggered that associates each of the now verified faces with a list of ten possible identities. This list of candidates is then disseminated for example through Twitter to a crowd of experts that vote and comment for their preferred identity.

Besides the identities of the different persons, all information that is associated to an image, such as the time or the place where the image was taken as well as contextual information about associated historical events can be reviewed by expert users and delegated to a crowd of domain experts for review.

Data aggregation, visualisation and analysis

Building on the computed co-occurrence of persons in images a social graph is constructed that connects them with each other. Connections gain in strength the more often persons appear together in an image. Finally the result of this process is depicted in a visualization of the social graph with a set of analytical tools.

The social graph in the History of Europe App aims at representing and visualizing dependencies between historically relevant persons in the context of European integration. Thereby the weight of the (social) links between person entities relies on their co-occurrence in historic photographs as identified by the aforementioned image indexation process. The more frequently two persons appear in different photographs, the stronger the link between the corresponding entities in the graph.

Users can interact with the History of Europe social graph in different ways, e.g. a click on a node results on an ego-graph of the selected person and clicking on an edge displays documents that relate to both selected relationship. As the documents stored in the collection very often come with a date of creation, the graph can be filtered by date with the timeline, displaying only the connections of documents created within this timespan. This timeline also shows the amount of photos per date that are contained in the collection. Another filtering option is the number of connecting documents, which allows the visualization of those relationships that are only included in an interval of a minimum and maximum number of documents. This feature is useful to highlight highest co-occurrences. Finally, the number of appearances of a person in the processed collection lets us identify people who appear particularly often in any given time frame.

Crowd discussion and a new approach to the representation of truth in digital research tools

Another challenge for the HoE app and the domain of the Digital Humanities in general is the conception of truth, which differs significantly e.g. to the conceptions of truth in Computer Science. Computer Scientists can rely on a stable foundation of what is true: Any experiment can be replicated and measured precisely. In the humanities the concept of truth is far more

complex: It is based on the insight, that there is no neutral or objective way to study human environments. The way, in which questions are asked, how data is selected to answer them, by what means this data is analyzed and finally the way in which the results of such analyses are communicated and received all challenge the idea of "one truth".

In order to represent the discursive nature of truth in the humanities within HoE we make use of a community-driven tool for question answering, similar to stackoverflow.com. User have the opportunity to answer questions and thus benefit from the knowledge within the expert crowd. However, the system allows for more than one answer and offers its users the possibility to vote and answer up or down, thereby allowing more than one answer to enter in competition with each other whilst also maintaining the full spectre of the discussion.

Summary and outlook

The History of Europe application takes on the challenge to combine cutting edge research in the domains of computer science, the design of human-computation tasks, data visualization, social engineering and the humanities by identifying synergies between the disciplines' strengths and by compensating for their weaknesses. We do this by building a pipeline which connects face recognition tools, data visualization and input from humans and creates an ongoing cycle of iteratively improved user input and machine output. The History of Europe application stands in line with a range of other online tools for historical research but introduces new social features as well as crowd sourcing from both click-workers and expert users which continuously improves the system. In the future we will expand the selection of sources to include digitized text documents as well as audio and video interviews from different archives.

Kanripo and Mandoku: Tools for git-based distributed repositories for premodern Chinese texts

Wittern, Christian
Kyoto University, JP
cwittern@gmail.com

Introduction

During the last 15 to 20 years, a considerable amount of premodern Chinese texts have been made available electronically, both for free and unhindered use and commercially in dedicated and locked down applications. Examples for the first type include projects such as the *Chinese Buddhist Electronic Text Association* (www.cbeta.org) and *Wikisource* (zh.wikisource.org) and the *Internet Archive* (www.archive.org), while examples for the latter includes products like the *Siku quanshu electronic edition* (四庫全書 電子版) by Digital Heritage Publishing or *Zhongguo jiben guji ku* 中国 基本古籍库 by Airusheng.

For scholars wanting to make use of these resources for their research, there are a few obstacles, including:

- different formats and ways to access the texts
- in many cases, texts do not conform to philological standards
- researchers can not annotate the texts and share their notes

Now, the projects described here attempt to develop an infrastructure for enabling scholars to work with repositories of freely available texts using a rapid prototyping approach with an expanding group of scholars for testing and early adoption. Technically, the main idea is to develop this as a network of repositories of texts, where each node a network consists of

a set of git repositories, that can represent multiple editions of texts.

Kanripo: A repository of premodern Chinese texts

First experiments for one node of such a network have been started at Kyoto University's Institute for Research in Humanities and its associated *Center for Informatics in East-Asian Studies*, CIEAS. In this experiment, the distributed version control system (DVCS) git is used as a basic transportation layer. Every text in the repository is represented by one DVCS node; different editions of the text can be represented by different versions or "branches" within this node; digital facsimiles can be associated with such versions. Users can also "fork" public projects and create new branches with their own annotations and comments and share these with other researchers, either in closed groups or with the general public.

The interaction with Kanripo occurs mainly through the web interface, but can also occur directly from the desktop tool Mandoku. However, as the experience so far has shown, it seems necessary to further develop the web interface, to enable it to become not just a hub for interaction between the users, but also a full-fledged client for editing texts in the repository or add comments and annotations. For this purpose, a system similar to the popular Github site (github.com) is envisioned, starting from a open source clone of github called gitlab. This is in a very early stage of development and any feedback from the audience will be much appreciated.

Mandoku: A tool for interacting with the repository

Development has also started aiming at a convenient desktop based tool for interacting with the repositories; a preliminary development version of this tool called *Mandoku* will also be demonstrated during the presentation.

Mandoku tries to meet researchers of premodern Chinese texts where they spent most of their time, that is reading, annotating and translating texts. This is why the current prototype is build on the powerful and extensible editor Emacs, while as a future implementation a interface for more casual users is also planned. To incite users to overcome the initial hurdle of adopting to a new and unusual editing program, a number of tools have been implemented, that enhance the usefulness of the system, among them a keyword in context (KWIC) index is generated on each text in every repository node, which can then be queried through Mandoku and the aggregated results will be displayed, a cumulative index for the query of dictionaries and specialized reference works; further text-analytic tools are planned.

Repository for digital publications

Another problem this system tries to address is a serious problem with the current mainstream form of digital publication: Currently a website serves usually as the main and in many cases as the sole venue of publication, thus usually hiding a complex textual resources between one browser-mediated interface (For a further discussion of this problem and a model for overcoming it see¹, {6}, {7} and {8}). This topic has been discussed for some time and valid suggestions and a discussion of the requirements can be found in², ³, ⁴ and⁵. Here, this proposal is taken up and expanded, namely by adding the requirement that the text will not only be made available to the scholarly community, but that it also be able to annotate it in a way that can be owned by the scholar adding the annotation and still be shared with interested colleagues.

In the framework presented here, fulfilling these requirements on a technical level is constructed as follows:

- A text available from the repository (which can be edited only by the editors) is "forked" into a private text repository on this or any other node in the network.

- The researcher can now edit and annotate the text to its hearts content, if so desired making the annotations available to others through pushing them to the forked repository.
- Occasionally, the researcher might come across errors in the text or has material he would like to offer for inclusion in the authoritative published text. He now issues a "merge request", which alerts the editors of the published texts to the existence of this piece of information, which usually is already visible on the private repository.
- The editor will consider the merit of the correction and can then incorporate it into the published text. When doing so, the origin of this information and any communication concerning the reason and argument for this correction will retain their relationship to this piece of text and is available as part of the scholarly record for this text.

If used correctly, this mechanism could provide a solution to the above mentioned problems with online digital publications. First experiments with scholars connected to the Kanripo research project showed that the technical protocol is not easily transparent to the scholars who are supposed to use it and does require more fine-tuning and better tool support in order to become more widely acceptable.

Conclusion

In an attempt to provide a stable, extensible platform for the curation of the textual heritage of China, a blueprint for text repositories that can form a network of related, but independent repositories for critically edited texts has been provided and a prototype of this implemented as the *Kanseki Repository of texts* (Kanripo), which can be accessed through one prototype client *Mandoku*. Further development will occur in collaborative form with scholars using this framework by attending to their emerging needs and thus hopefully developing into sustainable resources that can provide a solid base for all kinds of scholarly inquiry that relates to premodern Chinese texts.

References

1. **Wittern, Christian** (2013) "Beyond TEI: Returning the Text to the Reader". In: Journal of the Text Encoding Initiative [jtei.revues.org/691], 2013, 4.
2. **Robinson, P.** (2009). "Towards a Scholarly Editing System for the Next Decades". In: Sanskrit Computational Linguistics: First and Second International Symposia Rocquencourt, France, October 29-31, 2007 Providence, RI, USA, May 15-17, 2008 Revised Selected Papers. Springer London, Limited, pp. 346-357.
3. **Schmidt, Desmond and Robert Colomb** (June 2009). "A data structure for representing multi-version texts online". In: International Journal of Human-Computer Studies 67.6, pp. 497-514.
4. **Shillingsburg, Peter** (2010). "How Literary Works Exist: Implied, Represented, and Interpreted". In: Text and Genre in Reconstruction: Effects of Digitalization on Ideas, Behaviours, Products and Institutions. Ed. by Willard Mc- Carty. Cambridge: Open Book Publishers.
5. **Siemens, Ray et al.** (2009). "It May Change My Understanding of the Field: Understanding Reading Tools for Scholars and Professional Readers". In: DHQ: Digital Humanities Quarterly 3.4.
- Christian Wittern** "Towards an Architecture for Active Reading", in: Scholarly and Research Communication (www.src-online.ca/index.php/src/article/view/59), 2013, volume 3 / issue 4 p.1-11.
- Christian Wittern**, "The Digital Daozang Jiyao – How to get the edition into the Scholar's labs", in: digital humanities 2012. Conference Abstracts, Hamburg, 2012, p.422-424.
- Christian Wittern**, "Text Representation and Interchange in the Digital Age", at Annual conference of the Japanese Association for Digital Humanities 2012 at University of Tokyo, Sep. 15-17, 2012.

A Morphological Analysis of Classical Chinese Texts

Yasuoka, Koichi

yasuoka@kanji.zinbun.kyoto-u.ac.jp

Institute for Research in Humanities, Kyoto University, Japan

Yamazaki, Naoki

ymzknk@kansai-u.ac.jp

Faculty of Foreign Language Studies, Kansai University, Japan

Wittern, Christian

wittern@kanji.zinbun.kyoto-u.ac.jp

Institute for Research in Humanities, Kyoto University, Japan

Nikaido, Yoshihiro

nikaido@kansai-u.ac.jp

Faculty of Letters, Kansai University, Japan

Morioka, Tomohiko

tomo@kanji.zinbun.kyoto-u.ac.jp

Institute for Research in Humanities, Kyoto University, Japan

The most difficult point in the digital analysis of classical Chinese texts is that they don't have any spaces or punctuations between words or between sentences. They consist of continuous strings of Chinese characters from the start to the end of texts. Contrary to the analysis of modern Chinese texts, which have several punctuation marks and can be fragmented into phrases with these punctuation marks, the analysis of classical Chinese texts has to begin with finding out the ends of sentences.

Classical Chinese is an isolative language, which doesn't have any inflection or agglutination. Furthermore, we don't have any generally accepted word-class system for classical Chinese. We first ought to develop machine-supported word-class system for classical Chinese. However, in classical Chinese, many morphemes may be observed as nouns and verbs, etc. In this paper we propose a method to analyze classical Chinese texts. In our method, we use our original morphological analyzer based on MeCab¹. We propose a new four-level word-class system for classical Chinese on the MeCab-based analyzer. We design the top level of the word-class system to represent the predicate-object structure of classical Chinese. The second level is the ordinary word-class of classical Chinese. The third and fourth levels are word-subclasses to describe detailed behavior of the words in classical Chinese texts.

The development of our four-level word-class system for classical Chinese was not straightforward. At the early stage, we developed a prototype dictionary from IPA Japanese Dictionary² and defined a prototype word-class system for classical Chinese. We also developed a prototype corpus along the prototype word-class system. And then, at the later stage, we examined the prototype corpus and redefined our four-level word-class system to be more suitable and systematic for classical Chinese. Especially, we excluded "adjective" from the second level of our new word-class system, since, in classical Chinese, there exists no essential distinction between "verb" and "adjective"³. We refactored the prototype dictionary into our new dictionary, and the prototype corpus into our new corpus.

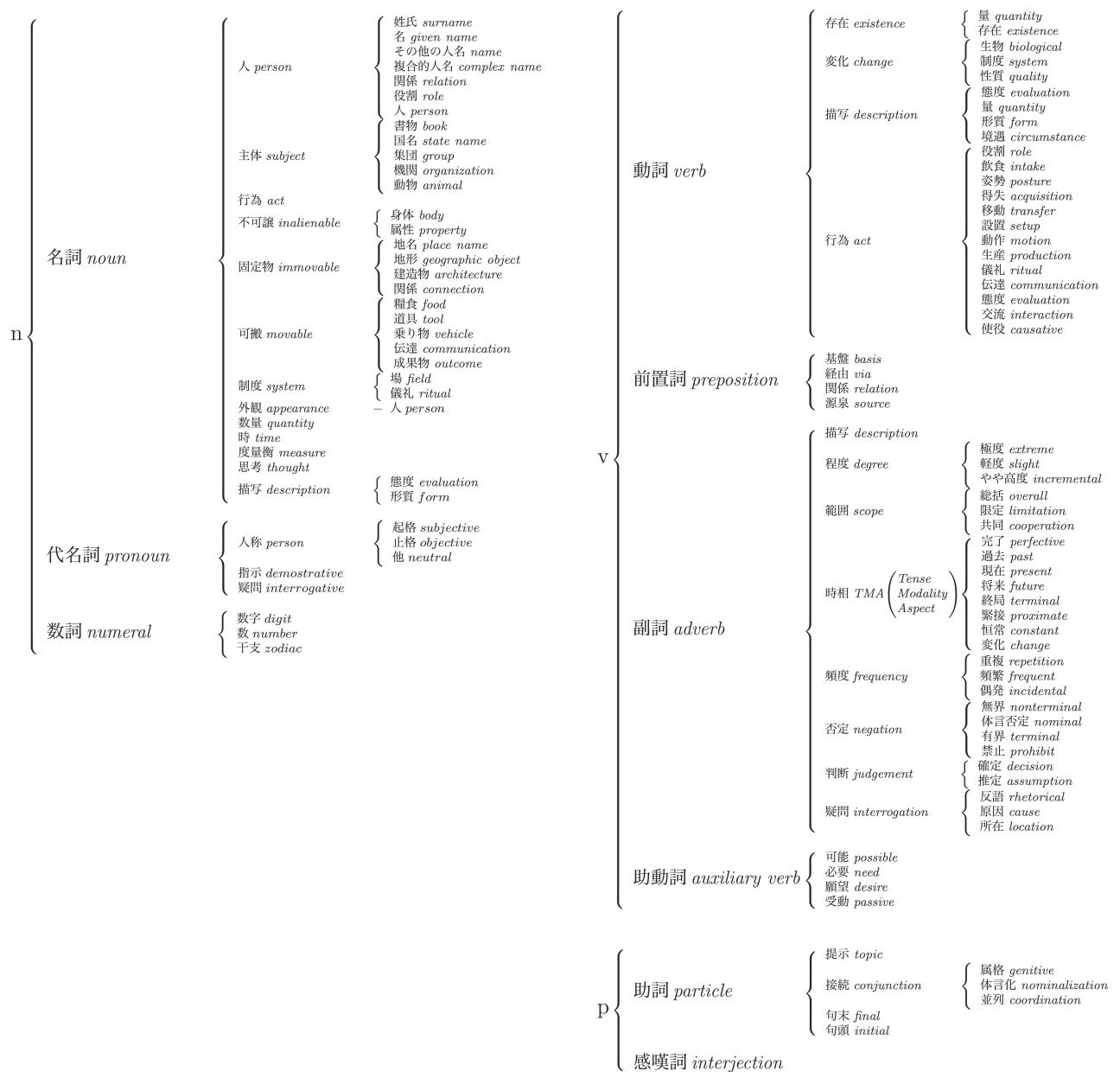


Fig. 1: Our Four-Level Word-Class System for Classical Chinese

In our new word-class system (Fig.1), the top level, which we call “word-superclass,” is defined to represent the predicate-object structure of classical Chinese: “n” represents objectives, “v” represents predicates, and “p” represents others. The second level is the ordinary word-class of classical Chinese: noun, pronoun, numeral, verb, preposition, adverb, auxiliary verb, particle, and interjection. We first constructed the word-class from a famous classical-Chinese dictionary Zenyaku Kanjikai⁴, and we reconstructed the word-class, especially excluding adjective. In our system, noun, pronoun, and numeral compose “n” word-superclass; verb, preposition, adverb, and auxiliary verb compose “v” word-superclass; particle and interjection compose “p” word-superclass.

The third and fourth levels are word-subclasses to describe detailed behavior of the words in classical Chinese texts. We first tried to construct these word-subclasses from *Word List by Semantic Principles*⁵. However, its levels were stratified too deep and its category was highly depended on Japanese. Therefore we constucted rather shallow word-subclasses, suitable for a morphological analysis of classical Chinese texts, from scratch (Fig.1). We have often revised the third and fourth levels of our word-class system. Whenever we revise our word-class system, we should modify our dictionary and corpus.

For the development of a large corpus, the collaboration of linguistic experts, scholars of classical Chinese, input

operators, and data managers is required. We use a distributed version control system, Git, to support the collaboration for the development of our corpus. Git is a powerful but complicated system, so we restrict our use of Git to avoid conflicts between versions of our corpus. And we have developed our own “skin” to hide the complicatedness of Git. Our own “skin” mainly consists of Git-based corpus manager, our Mecab-corpus editor (mentioned below), a system updater of our dictionary and corpus, and a system updater of the framework.

In order to make corpus for classical Chinese on MeCab, we have constructed a MeCab-corpus editor based on XEmacs CHISE⁶. We use the MeCab-corpus editor to compile our digital corpus and our digital dictionary based on our four-level word-class system for classical Chinese (Fig.2). In our MeCab-corpus editor we first input typical sentences from classical Chinese texts. Second we push the right-most button “classical Chinese” of the editor, then we obtain a morpheme sequence temporarily segmented by MeCab. Third we edit the sequence to categorize its words, looking up authoritative textbook references of the sequences. And last we include the morpheme sequence in our corpus for classical Chinese.

Our corpus for classical Chinese on MeCab now includes about 20,000 sentences, written in our four-level word-class system. Our dictionary for classical Chinese on MeCab includes about 5,000 words, which we categorized into our four-level word-class system. We keep increasing our corpus, and we also keep selecting new words from our corpus to add them into our dictionary.

In conclusion, we made a morphological analyzer for classical Chinese. The analyzer required a dictionary and a corpus based on a word-class system. We developed our four-level word-class system, suitable for analysis of classical Chinese, originally made from some other dictionaries, and then we reconstructed the word-class system. We also developed the Git-based framework including our Mecab-corpus editor, which allowed us to edit the corpus and dictionary effectively.

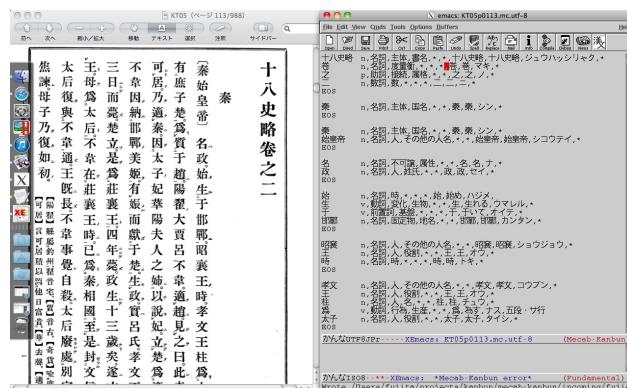


Fig. 2: Screenshot of an Authoritative Textbook and Our MeCab-Corpus Editor

References

1. T. Kudo, K. Yamamoto and Y. Matsumoto (2004): *Applying Conditional Random Fields to Japanese Morphological Analysis*, Conference on Empirical Methods in Natural Language Processing, pp.230-237.
2. [mecab-ipadic, code.google.com/p/mecab/downloads/detail?name=mecab-ipadic-2.7.0-20070801.tar.gz](http://mecab-ipadic.googlecode.com/p/mecab/downloads/detail?name=mecab-ipadic-2.7.0-20070801.tar.gz)
3. N. Yamazaki, T. Morioka and K. Yasuoka: *Refactoring of Wordclasses for Morphological Analysis of Classical Chinese*, The Computers and the Humanities Symposium 2012, pp.39-46.
4. Y. Togawa, et al. (2011): *Zenyaku Kanjikai*, 3rd Ed., Sanseido.
5. National Institute for Japanese Language and Linguistics (2004): *Word List by Semantic Principles*, Revised & Enlarged Ed., Dainippon Tosho.

6. T. Morioka (2008): *CHISE: Character Processing Based on Character Ontology*, 3rd International Conference on Large-Scale Knowledge Resources LKR, pp.148-162.

Collaboratively maximizing inter-ontology agreement for controversial domains: A case study of Jewish cultural heritage

Zhitomirsky-Geffet, Maayan

maayan.geffet@gmail.com

Information Science Dept., Bar-Ilan University

Erez, Eden Shalom

tempeden@walla.com

Computer Science Dept., Bar-Ilan University

Ontology is a semantic scheme which comprises the main classes (concepts) of the given domain of knowledge, their properties, inter-relationships and instances¹. The basic ontological relationship types are IS-A (hyponymy) and instance-of (a concrete type-of) which are the basis of the conceptual taxonomy, and other thesauri-like (e.g. "part-of", "related-to") or content-driven relationships (e.g. "located-at", "produced-by") are allowed as well. One of the standards for formal encoding of ontologies is RDF (Resource Description Framework)². RDF-based knowledge representation consists of triples (relations) of the form: (concept1 – relationship type – concept2). For example, in the domain of Jewish cultural heritage the following relations could be included in the ontology: (Passover IS-A Jewish Holiday), (Holiday part-of Cultural Heritage), (Orthodox Jew preserves Jewish Tradition). Every triple corresponds to a certain statement or fact on the domain.

Nowadays, ontologies are widely used as a formal domain vocabulary for content-specific agreements in a variety of knowledge-sharing activities, such as, information organization, retrieval and tagging³. However, in the current state of the Semantic Web for many domains there are multiple diverse ontologies rather than one standard vocabulary. This is due to the fact that they are typically constructed by different experts who often possess contrary viewpoints especially in cases of controversial domains. These domains include cultural heritage, economy, politics, history, religion, art and even medicine. Apparently, when using several ontologies in a common application, mismatches can create incoherent results for the users. Therefore, reaching maximal agreement between these ontologies is necessary to standardize and unify the domain vocabulary. Hence, building unified consensual ontologies has become a big research challenge.

The objective of this work is to explore ways to maximize the inter-ontology agreement for controversial domains. Particularly, we experimented with the case of the Jewish life style domain which comprises cultural, religious and political aspects. We also aim to explore whether it is possible to identify consensual ontological relations from diverse ontologies and construct a maximal subset of consensual vocabulary for the controversial domain.

Research Methodology

To overcome the semantic heterogeneity problem in ontologies a variety of ontology matching algorithms were presented in the past decade⁴,⁵. These systems usually focus on mapping individual concepts (and/or their taxonomic structures) of one ontology to similar concepts in the other one. However, the level of inter-ontology agreement assessed by the automated approaches is limited by the following major factors:

1. The algorithm's ability to recognize semantically similar concepts, which are frequently conveyed by different terms;

2. Matching of isolated concepts which does not reveal the maximal potential for semantic similarity of ontologies unless all the direct and indirect relations (triples) binding these concepts can be consistently matched as well.
3. The low overlap between the explicit terminologies of diverse ontologies (for both, concepts and relations) due to the viewpoint diversity of their composers (especially for controversial domains).

Some partial solutions were lately proposed in the literature. Thus, in order to reduce the impact of the first factor, a recent study by⁶ proposed to employ "wisdom of crowds" for detecting similar concepts in two different ontologies. To resolve the second limitation, similarity between ontologies should be computed for relations rather than for individual concepts, as implemented by⁷. He counted exactly and partially matching triples from a pair of given ontologies. This methodology is adopted in the current research which further focuses on matching relations rather than individual concepts. However, the third and the most crucial factor, a relatively small amount of common relations, still remains unresolved.

The main question is how to reveal and assess the potential maximal agreement between ontologies despite the low overlap between them. The essence of this problem is the underlying assumption that relations, which are present in one ontology, but are missing from the other ontology, are automatically considered as unmatched and increase the ontology disagreement level. Nevertheless, it can be observed that if the ontology composer did not choose to add a relation to his personal ontology, it is unclear whether he agrees or disagrees with the truth of this relation. To this end, we introduce a new collaborative approach where independent ontology composers can explicitly express their opinions on the others' relations. Thus, after completing the construction of their own ontologies the participants are exposed to the relations of the others and are asked to decide for each of them whether it is true or false. Then, the "real" exhaustive inter-ontology agreement can be calculated based on these votes rather than by counting the common relations in the original ontologies.

We distinguish between two levels of ontology agreement:

- The *local* agreement between a pair of ontologies that can be calculated as follows:

$$LA = \frac{CR + AO_1 + AO_2}{O_1 + O_2 - CR}$$

CR- number of common relations in two ontologies

*AO*₁- number of relations in the first ontology that a composer C2 agrees with

*AO*₂- number of relations in the second ontology that a composer C1 agrees with

*O*₁- number of relations in the first ontology

*O*₂- number of relations in the second ontology

Fig. 1: The local agreement measure

- The *global* agreement definition between all the ontologies for the domain:

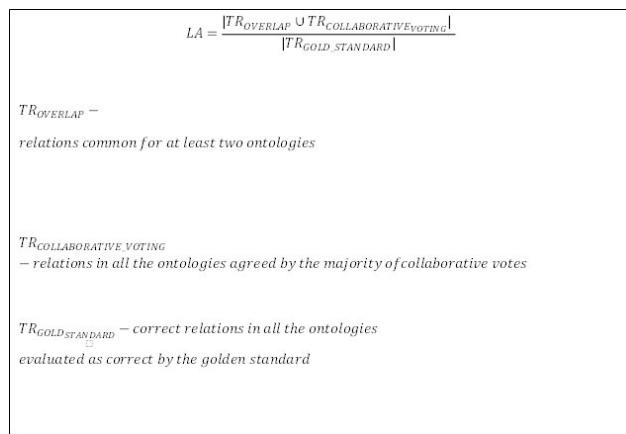


Fig. 2: The global agreement measure

These relations constitute the consensual part of the ontologies. The other relations will be considered as controversial. The threshold discriminating the consensual and controversial relations can be computed by applying machine classification on the composer's votes as features for each relation as described in the next section.

Experimental Setting and Results

Based on the above collaborative scheme for relation evaluation we conducted an experiment with 21 ontology composers (students of the Semantic web course in Information Science Department). At the first step, the group has chosen a set of 130 concepts which are the most representative of the domain of Jewish life style. Then, every participant was required to construct up to 100 RDF-style triples (relations) with the above concepts and a set of 15 predefined relationships (such as, IS-A, instance-of, part-of, disjoint-with, entails, located-at, antonym-of) independently from the other members of the group. The relations were inserted into the web-based system implemented for this purpose. Further, each one of 1175 distinct ontological relations, created by all the participants at the first step, was consecutively displayed by the system and independently evaluated as true or false by every participant of the group. The analysis of the results shows that the initial local agreement between the diverse ontologies (the *CR component of the measure*) was very low (0-22%) reflecting the controversy of the domain. This is despite the fact that all the participants used the similar set of concepts and relationships for relation construction. The exhaustive local inter-ontology agreement assessed after applying the collaborative evaluation procedure appears to be much higher (39-90%), as demonstrated in Fig. 3. Thus, our collaborative scheme substantially enhances the local agreement level between ontologies.

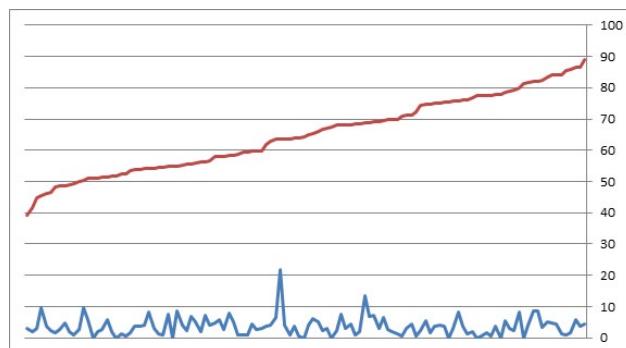


Fig. 3: The local agreement rates computed by the amount of overlapping relations in the original ontologies as a baseline (the blue skew) vs. by the votes after applying the collaborative evaluation

procedure (the red skew). Axis X represents pairs of different ontologies in our corpus.

To create a golden standard evaluation required for the global agreement calculation, two experts were asked to annotate all the statements as correct (ground truth) or controversial (depending on one's personal beliefs). First, they worked independently and then reached full consensus through a discussion. Overall, 885 (out of 1175) relations were judged as correct facts (TRgoldstandard), and 290 as controversial viewpoints. For example, the correct relations (like, Passover – is-a – Jewish holiday; Reform Jew – disjoint-with – Orthodox Jew; Western Wall – located-in – Jerusalem) are expected to obtain the large majority of agreement ("true") votes in the collaborative procedure, while the controversial relations (such as, Ultra-Orthodox Jew – resists – Scientific progress; God – created – Universe; Bible – written by – Man) are supposed to gain intermediate scores for both agreement ("true") and disagreement ("false") voting categories.

The $|T_{Overlap}|$ component needed for the global agreement calculation was rather low 29% (232 out of 885) even with the threshold of 2, as only 256 relations appeared in at least two ontologies. 919 (almost 80%) of the relations appeared in one out of 21 ontologies), while only one relation was present in 9 out of 21 ontologies.

Then, in order to estimate the global agreement after applying the collaborative evaluation procedure, we utilized the WEKA environment⁸ to choose the optimal machine classification algorithm. Eventually, the best 10-cross validation results were achieved by the Multilayer Perceptron algorithm which yielded 90% average accuracy. As a result $|T_{collaborativevoting}|$ of 876 was obtained. Interestingly, all 232 relations of $T_{Overlap}$ were included in $T_{collaborativevoting}$. Overall, 99% of the correct relations according to the golden standard were classified as correct by the automatic classifier. The classifier used as features the true and false voting scores. Most of the errors in classification were controversial relations probably reflecting some common viewpoint among the members of the group. So in the future research we intend to conduct crowdsourcing microtask-based experiment (like in⁹) with a much larger number of participants.

In summary, our collaborative method significantly increases the baseline agreement that can be achieved manually or automatically from the explicitly overlapping/matching relations. This methodology further leads to construction of a reliable large consensual ontology for controversial domains which seem impossible to achieve from the small overlap of the original ontologies.

References

1. Noy, N.F. & McGuinness, D.L. (2001). *Ontology Development 101: A Guide to Creating Your First Ontology*. Technical Report KSL-01-05, Stanford Knowledge Systems Laboratory and Stanford Medical Informatics Technical Report SMI-2001-0880, Publisher: Citeseer, (pp. 1-25) doi: 10.1.1.136.5085
2. Hayes, P. and B. McBride. RDF Semantics. W3C Recommendation: www.w3.org/TR/2004/REC-rdf-mt-20040210/ (2004 last accessed July 2013).
3. Gruber, T. R. (1993). *Toward principles for the design of ontologies used for knowledge sharing*. International Journal of Human-Computer Studies, 43(5-6), 907-928. Retrieved from itee.uq.edu.au/~infs3101/_Readings/OntoEng.pdf
4. Shvaiko P. and J. Euzenat (2013). *Ontology matching: state of the art and future challenges*. IEEE Transactions on Knowledge and Data Engineering, 25(1): 158-176.
5. Flouris G., Manakanatas D., Kondylakis H., Plexousakis D. and G. Antoniou (2008). *Ontology change: classification and survey*. The Knowledge Engineering Review, 23(2), 117-152.
6. Sarasua, C., Simperl, E. and N.F. Noy. (2012) *Crowdmap: Crowdsourcing ontology alignment with microtasks*. Proceedings of the International Semantic Web Conference, Boston, USA, pp. 525- 541.

7. d'Aquin, M. (2010). *Formally measuring agreement and disagreement in ontologies*. Proceedings of the Fifth International Conference on Knowledge Capture, Redondo Beach, California, USA, pp. 145-152.

8. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and I. H. Witten (2009). *The WEKA Data Mining Software: An Update*. SIGKDD Explorations. 11(1): 10-18.

9. Sarasua, C., Simperl, E. and N.F. Noy. (2012) *Crowdmap: Crowdsourcing ontology alignment with microtasks*. Proceedings of the International Semantic Web Conference, Boston, USA, pp. 525- 541.

The Changing Canon of Beauty: Facial Attractiveness in the Representation of Human Faces in World Painting

de la Rosa, Javier

versae@gmail.com
CulturePlex Lab, Western University

Caldas, Natalia

ncaldas@uwo.ca
CulturePlex Lab, Western University

Dutta, Nandita

ndutta2@uwo.ca
CulturePlex Lab, Western University

Suárez, Juan Luis

jsuarez@uwo.ca
CulturePlex Lab, Western University

1. Introduction

The face and its proportions have always captured our attention and produced fascination. Even newborns have been reported to dedicate more time to attractive faces than others¹. How these proportions are meant to be the guidelines that define facial beauty, has been an object of study since the times of Plato. However, absolute approaches, such as Hogarth's serpentine line², the Vitruvius' "well-shaped man"³, *divina proportiones*, or mathematically based ratios such as golden, Fibonacci, or 1:1.6, have proven insufficient to explain how beauty actually works⁴. As Francis Galton said⁵, "The general expression of a face is the sum of a multitude of small details, [...]." We can now afford to extend this concept and say that the attractiveness of a face is also the sum of a varied set of distinct features. Recent research on evolutionary psychology and neuroaesthetics suggest the same. Beauty of unknown faces seems to include aspects from averageness, symmetry, sexual dimorphism, pleasant expressions, and youthfulness^{6 7}^{8 9}.

In this study we take state-of-the-art research results on attractiveness and beauty and extrapolate them to the analysis of faces in world painting. We first collected a data set of over 120,000 paintings, and applied industry standard face recognition algorithms to extract facial traits. Furthermore, based on meta-analysis of symmetry and averageness¹⁰, we established clues on whether faces across time could be considered more beautiful, and when these trends occur.

2. Method

2.1 Data-set

The data-set was obtained from pintura.aut.org, a nonprofit organization working on autism¹¹. The whole set of 120,000

images of paintings was narrowed down to 25,000 images by removing the paintings with no faces. The number of faces found is about 47,000. The distribution of the number of faces per paintings follows a power-law that fits into the Pareto principle.

For the current study only 5,800 faces were taken into account: frontal faces no smaller than 150 pixels in height, with pitch and yaw angles between 10° and -10° with respect to the vertical line, and with valid information for at least the following traits: eyes, nose, mouth, height, width, and center of the face. Face rotation or roll was fixed geometrically. Our analysis covers the period between 13th and 19th centuries.

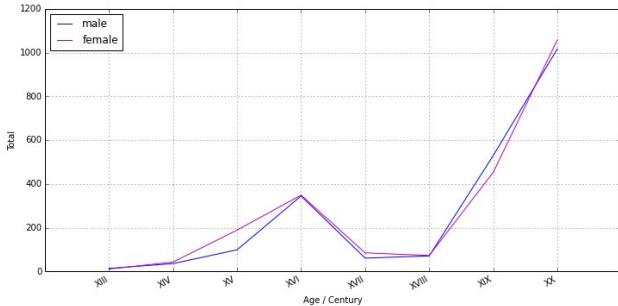


Fig. 1: Distribution of the number of male and female faces per century

2.1. Averageness

For each century an average male and female face has been computer-generated (see Figure 2), in addition to a non-gender-specific one that combines both genders. In order to produce this averaged composite face, we first centered the faces according to the *center* point given by the face recognition algorithm. Faces were then resized to make them fit into a PNG canvas of 500 by 500 pixels at 300dpi resolution, and given a height of 200 pixels (faces with height lower than 150 pixels were excluded to avoid blurred pixelation of the average face). This process was achieved by using affine and projective 2D transformations¹² from the original painting to the desired canvas. Every face standardized by size was converted into a 3D numerical matrix representing each of the layers of the RGB color model. A regular statistical mean was then calculated over the set of faces of each century in order to obtain the average value for each pixel. Once the average matrix was calculated, it was converted back into a PNG image.

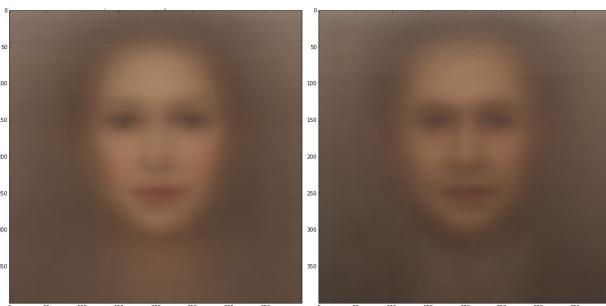


Fig. 2: Average 20th Century female face from 1058 faces (left), and male from 1017 faces (right).

Resulting quality and averageness of the composite rely on the number of faces used in each century for generating the averaged face. The same face recognition algorithm used in the data-set is then applied to averaged composites. This allows us to measure the averageness of a face as the difference between its symmetry and the symmetry of the average face for that particular period. Averageness refers to the degree to which a given face resembles the majority of faces. In our study averageness values go from the most average, 0, to the least, 1.

2.2 Symmetry

Calculation of symmetry is commonly based on Grammer and Thornhill's early work¹³. Their method makes use of 12 different points (one more for averageness): 2 for each eye, 2 for the nose, 2 for the mouth, 2 for the cheekbones, and the last 2 for the jaw. With those, they create lines for each pair and calculate their midpoints. In a perfect symmetrical face, all midpoints must lie on the same vertical line. Although we did not work with those 12 points, our algorithm still used 3 points for the mouth (left, center, and right), 1 for each pupil, and 1 for the nose. This number of traits proved enough for symmetry calculation; therefore, even though our methodology is slightly different from the one proposed by Grammer and Thornhill, the main idea remains unchanged.

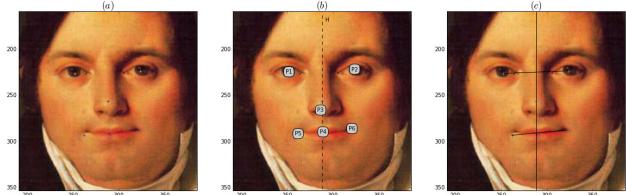


Fig. 3: (a) Example of face and detected points for eyes, nose, mouth and center. (b) Vertical line, H, to divide the face into two hemifaces, and numerated points for all the features. (c) Lines for calculating distances between midpoints and hemiface line.

Additionally our algorithm also gave us the centroid or geometric center of all detected features (Figure 3a), which can be taken as the center of the face. From it, we can set a straight line that splits the face into two sides or hemifaces. Figure 3b shows points 1 to 6 (P1 for left eye, P2 for right eye, P3 for nose, P4 for mouth center, P5 for left mouth corner, and P6 for right mouth corner), as well as the line H, that we assume to be the axis of face symmetry. We now trace segments: D1 between P1 and P2, and D2 between P5 and P6 (Figure 3c). For these segments we calculate the midpoints M1 and M2. Symmetry is now obtained as the sum of the distances in pixels of M1, M2, P3 and P6 with respect to the line H. Only lateral symmetry is therefore estimated. For perfect symmetrical faces this value adds to zero; all symmetry values are normalized between 0 and 1.

3. Results

Figures 4 and 5 summarize the averages calculated per century for averageness and symmetry values, respectively.

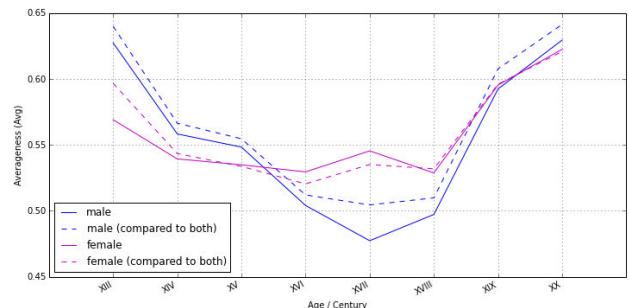


Fig. 4: Distribution of average values of averageness per century for female and male faces conforms to their own gender-specific averageness (solid lines), and to the combined averageness with both genders (dashed lines).

Figure 4 shows the distribution of averageness for male and female faces compared to their gender specific averaged composites. In dashed lines we can also see the same distribution but in regards to non-gender-specific average face. A quick two-sample Kolmogorov-Smirnov test allows us to see that there is no significant difference between the two male distributions ($p=0.92$) and the two female ones ($p=0.51$).

For male faces, we observe that the levels of averageness are low in the 13th Century, but then begin to decrease until the 17th Century, which leads to a gradual increase until the 20th Century. Averageness, difference between faces and the averaged composite face of each century, can give clues on how similar faces are to each other. Therefore, the peak seen in the 17th Century may be explained by the recovery of the Greek style in Neoclassicism where the faces depicted were following the same pattern, resulting in a closer distance for each one to the average face.

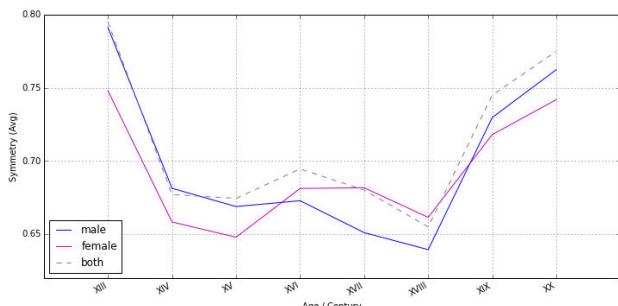


Fig. 5: Distribution of average values of symmetry per century for male faces (blue), female (magenta), and both (dashed gray).

Average values of symmetry per century are shown in Figure 5 for male, female, and both. We see that most symmetrical female faces were found in the 15th Century and male faces in the 18th Century. After that, all faces rapidly become asymmetrical during the 19th and 20th Century. This might be explained by the then new art styles, such as Rococo, that rejected the concept of symmetry from previous styles, such as Baroque and Neoclassicism.

4. Discussion and Further Research

As reported by previous studies (e.g.), there might be a link between averageness and attractiveness evaluations, which suggests that the more average the face, the more attractive it is perceived to be. Therefore, representations closer to the mean tendency of a population are preferred rather symmetry. Extrapolating these ideas to our results (see Figure 4): male faces were more attractive in the 17th Century; unlike female faces, that experienced a decrease in averageness for the same period, hence, so did their perceived beauty.

Results from the symmetry analysis seem to support the same trend: faces were more symmetrical, and allegedly more attractive, between the 14th and 17th Century. These results would be further supported by means of rating experiments. We have found differences between genders that merit more research.

Finally, while the purpose of this study was to establish when faces were seen as more attractive, some researchers have noticed a weak link between beauty and health that should be explored in the future. Skin tone in relation to attractiveness is also a topic that seems to be gaining more interest in the last years.

References

1. Grammer, Karl, and Randy Thornhill (1994). *Human («Homo sapiens») facial attractiveness and sexual selection: The role of symmetry and averageness*. Journal of comparative psychology 108.3: 233.
2. Hogarth, William (1772). *The analysis of beauty*. Printed by W. Strahan for Mrs. Hogarth.
3. Pollio, Vitruvius (1867). *De architectura*. Teubner.
4. Etcoff, Nancy (2011). *Survival of the prettiest: The science of beauty*. Random House Digital, Inc..
5. Etcoff, Nancy (2011). *Survival of the prettiest: The science of beauty*. Random House Digital, Inc..
6. Thornhill, Randy, and Steven W. Gangestad (1999). *Facial attractiveness*. Trends in cognitive sciences 3.12: 452-460.
7. Rhodes, Gillian, and Leslie A. Zebowitz (2002). *Facial attractiveness: Evolutionary, cognitive, and social perspectives*. Vol. 1. Ablex Publishing Corporation.
8. Berry, Diane S. (2000). *Attractiveness, attraction, and sexual selection: Evolutionary perspectives on the form and function of physical attractiveness*. Advances in experimental social psychology 32: 273-342.
9. Etcoff, Nancy (2011). *Survival of the prettiest: The science of beauty*. Random House Digital, Inc..
10. Rhodes, Gillian (2006). *The evolutionary psychology of facial beauty*. Annu. Rev. Psychol. 57: 199-226.
11. pintura.aut.org/ Accessed on Oct 31st, 2013.
12. Schneider, Philip, and David H. Eberly (2002). *Geometric tools for computer graphics*. Morgan Kaufmann.
13. Grammer, Karl, and Randy Thornhill (1994). *Human («Homo sapiens») facial attractiveness and sexual selection: The role of symmetry and averageness*. Journal of comparative psychology 108.3: 233.
14. Gombrich, Ernst Hans, and Ernst Hans Gombrich (1995). *The story of art*. Vol. 15. London: Phaidon.
15. Rhodes, Gillian (2006). *The evolutionary psychology of facial beauty*. Annu. Rev. Psychol. 57: 199-226.
16. Komori, Masashi, Satoru Kawamura, and Shigekazu Ishihara (2009). *Averageness or symmetry: which is more important for facial attractiveness?*. Acta psychologica 131.2: 136-142.
17. Wade, T. Joel (1996). *The relationships between skin color and self-perceived global, physical, and sexual attractiveness, and self-esteem for African Americans*. Journal of Black Psychology 22.3: 358-373.
18. Hill, Mark E. (2002). *Skin color and the perception of attractiveness among African Americans: Does gender make a difference?*. Social Psychology Quarterly: 77-91.
19. Swami, Viren, Adrian Furnham, and Kiran Joshi (2008). *The influence of skin tone, hair length, and hair colour on ratings of women's physical attractiveness, health and fertility*. Scandinavian Journal of Psychology 49.5: 429-437.

ClipNotes: Digital Annotation and DataMining for Film & Television Analysis

deWaal, Andrew

University of California, Los Angeles, United States of America

In the age of Big Data and technologyassisted research, scholarship in the humanities is developing innovative new approaches and methodologies, be they quantifiable and visual (Moretti, 2005), algorithmic (Ramsay, 2011), built upon 'distant' and 'machine' readings (Hayles, 2012) and 'cultural analytics' (Manovich, 2009), or of the many possibilities and techniques offered in the field of Digital Humanities (Burdick et al, 2012). Film and television analysis, however, has been slow to adapt digital datadriven research. This presentation will demonstrate a software project entitled ClipNotes, a software application created by Dr. Stephen Mamber at UCLA along with a team of graduate students that are helping to develop it. Examples will be shown of both traditional textual analysis amplified by the software and my own research that quantifies product placement and brand integration in film and television.

Though cinema and media studies has rightfully continued on its course of splintering and diversifying into a multitude of interdisciplinary amalgamations and subdisciplines, close textual analysis remains at the heart of what we do, particularly with regards to teaching. Technology has dramatically increased our ability to deconstruct a film text, from VHS to DVD to digital interfaces which allow even more minute control over image and sound. Creating clips and screengrabs is becoming easier and is an increasingly common feature of lectures and presentations, but the method of visual analysis

itself hasn't advanced very much, nor has the incorporation of many other research technologies. Notably, the use of programming and algorithmic analysis is quite limited in film studies research compared to other disciplines within the humanities which are utilizing digital tools to invigorate their methods and broaden their scope.

ClipNotes is an iPad and Windows 8 app that facilitates quick segmentation, annotation and presentation of film clips, an example of the research possibilities that are provided in a collaborative, database-driven, XML-based software environment. The design of ClipNotes is deceptively simple: it allows users to mark up video files with metadata and present this analysis. Start/stop times, clip descriptions, and captions are assembled in easily produced XML files, which stands for extensible markup language that is both human and machine readable. When the XML file is linked to the video file, precise, granular analysis is made possible and is easily presented and disseminated. A public repository for these XML files is available at clipnotes.org, which will allow for widespread sharing of textual analysis and should provide an invaluable teaching resource. Users are encouraged to upload their own XML files for inclusion and we have begun building an extensive database of freely available teaching and research materials. For obvious copyright reasons, the films are not included, but guides are available to demonstrate how DVDs can be easily and legally encoded into digital files under fair use exemptions. Then, the video is linked together with the XML file by the application. In practice, ClipNotes collapses the research and presentation process by bridging the two: your textual research and analysis is the development of your presentation.

To get the archive started, a series of films have been coded in XML and are available for teaching usage. *Citizen Kane*, of course, has been catalogued, providing quick access to its landmark visual style. All instances of deep focus, triangular framings, door and window framings, graphic matches, media representations, and the motifs of light and bulbs, bars and fences, and mirrors are quickly and easily accessible through ClipNotes. The ability to quickly but briefly demonstrate a series of scenes particularly ones involving sound or camera and character movement is a tremendous resource in a lecture or presentation situation. One of the distinct strengths of film is the symphonic arrangement of audiovisual patterns, something that is lost in merely showing screen grabs or a few extended scenes.

Beyond this kind of instructional usage, however, the practice of granular analysis can reveal new discoveries in even the most wellworn films. Hitchcock is likely the most thoroughly researched auteur, but with the ability to isolate and compare shots on a framebyframe basis, very subtle visual patterns can be identified. For instance, the ability to document the most minute patterns in *Psycho* reveals a microdetailed facial dramaturgy that adds significant visual emphasis to Hitchcock's already complex exploration of identity.

Douglas Sirk is another master of this kind of over-determined mise-en-scene, and a fitting subject for ClipNotes, as well as my own work that concerns the political economy of media as a textual phenomenon. *Written on the Wind* is a fine example, and for my analysis I catalogued over 100 examples of 4 distinct patterns: frame within frames and obscured frames, the mirror motif, vertical objects as phallic imagery, and what I hope is a new addition to the already extensive Sirk Studies archive: the product shot. While Sirk's satirical savvy is well-documented by scholars, the contradictory impulse the ways in which he helped promote and fetishize the very consumer culture he was satirizing is a less prominent element of his legacy. Barbara Klinger provides a useful corrective of his body of work in *Melodrama and Meaning: History, Culture, and the Films of Douglas Sirk*, in which she considers the various extra-textual discourses (academic, industrial, trade, popular press, star, gossip, camp) that have contributed to Sirk's legacy, but she expressly decides not to engage in any textual analysis, as her interest is in the surrounding discourse. I believe textual analysis could be a useful addition to furthering her nuanced reading of Sirk, particularly the industrial and promotional elements of his work that get overlooked in favour of his more critical characteristics.

Klinger shows how *Written on the Wind* was heavily promoted in women's magazines as a part of lifestyle marketing, particular tie-ins with its fashion, make-up and home decor. Textual analysis of the film itself reveals a similar impetus for commercial promotion, catalogued here as the "product shot." At least two dozen shots of consumer products are featured in the film, most prominently luxury cars, fashion, and jewellery. The first scene immediately establishes the patterns of framing, mirrors and vertical objects, but it also quickly pronounces the importance of products, by featuring Kyle's shiny, yellow sportscar. Far from claiming that this kind of quasiproduct placement deems Sirk some sort of sell-out or salesman, I think a more complete picture of all that Sirk accomplished in his films just adds to his stature. That he was able to satisfy the high standards of critics, scholars, studio executives, and promotional agents is a testament to the complexity of his filmmaking ability. It also provides some insight into the early history of product placement, brand integration, and transmedia.

ClipNotes and XML--encoding provides the opportunity to generate and catalogue large data sets of analytic material for audiovisual texts, lending quantification and data processing possibilities in the future. Sound and image are inherently more difficult to catalogue and quantify than the written word, which accounts for some of the delay in the use of digital humanities methods in cinema and media studies, but creative new digital tools should be able to bridge this gap. Similar to the Text Encoding Initiative, the datamining prospects generated by ClipNotes are impressive. With the assistance of digital tools, we will be able to both dig deep into solitary texts, discovering and quantifying micro-relationships, while also mapping broad, macro-cultural dynamics as a result of this wide-ranging data. In a digital era marked by the vast proliferation of complex texts, software applications can be utilized to enact a more rigorously detailed analysis, to archive and disseminate provocative insights, and to extend digital scholarship.

Thank you, and please visit clipnotes.org for more information.

References

- Burdick, Anne et al.** *Digital_Humanities*. The MIT Press, 2012.
- Hayles, N. Katherine.** *How We Think: Digital Media and Contemporary Technogenesis*. University Of Chicago Press, 2012.
- Klinger, Barbara.** *Melodrama and Meaning : History, Culture, and the Films of Douglas Sirk*. Bloomington: Indiana University Press, 1994.
- Manovich, Lev.** "Cultural analytics: Visualizing cultural patterns in the era of more media." *Domus*: 923 (2009).
- Moretti, Franco.** *Graphs, Maps, Trees: Abstract Models For A Literary History*. Verso, 2005.
- Ramsay, Stephen.** *Reading Machines: Toward an Algorithmic Criticism*. University of Illinois Press, 2011.

Distributed “Forms of Attention”: eMOP and the Cobre Tool

duPlessis, Anton Raymund

Texas A&M University, United States of America

Mandell, Laura

Texas A&M University, United States of America

Creel, James

Texas A&M University, United States of America

Maslov, Alexey

Texas A&M University, United States of America

A recent article by Paul Gooding, Melissa Terras, and Claire Warwick argues that a gap in our knowledge about the impact upon scholars of "large-scale digitized collections" of textual

data has spawned "myths" about mass-digitization—those surrounding the distant- vs. close-reading debate, as well as dystopian arguments about digitally disrupted attention spans (Gooding et. al.; Moretti; Trumpler; Guillory; Hayles). "Where understanding lags behind innovation," Gooding, Terras, and Warwick argue persuasively, "the rhetoric of technological determinism can fill the void" (633). Central to the myth that digital media produce "crowds of quick and sloppy readers" (632) is ignorance: the last decade has witnessed the emergence of a spate of digital editing and annotating tools as well as the emergence of what might be called "network editing."^[2] As has been amply revealed by the Australian Newspaper Digitisation Program, as well as other projects involving the crowd in correcting textual transcriptions such as Transcribe Bentham, people are as much engaged by the task of modifying digital textual archives as they are in using them (Halley, Terras), and textual "modding"^[3] by networks of users requires paying close attention to texts, as might networked monasteries of monks in the process of transcribing them. The need for distributed networks of people helping to solve problems endemic to creating large textual corpora, in other words, fosters close attention to text. This fact is demonstrated by the Early Modern OCR Project (eMOP)^[4]: the attempt to produce searchable text for 45 million page-images of texts published between 1473 and 1800 would fail were it not for the project's adaptation of Cobre, a tool originally designed for closely examining various 16th Century Iberoamerican imprints, and Cobre elicits careful scholarly attention from globally distributed experts and citizen scholars wishing to take part in improving the quality and kind of information available on the Internet.

The eMOP project focuses on how to digitize the archive of early modern texts, despite problems entailed. The printing process in the hand-press period (1473-1800), while systematized to a certain extent, nonetheless produced texts with fluctuating baselines, mixed fonts, and varied concentrations of ink (among many other variables). Adding to these factors, the quality of the digital images of these books is very poor: ProQuest's Early English Books Online dataset (EEBO) and Gale's Eighteenth-Century Collections Online (ECCO) contain page images that were digitized in the 1990s from microfilm created in the 1970s and 80s. Hand-press printing as well as skewed low-quality images with no gray-scale originals creates a problem for Optical Character Recognition (OCR) software. OCR engines are notoriously bad at translating into texts digital images of early modern texts even under the best of circumstances (Gooding, Terras, and Warwick). That is trying to translate the images of these pages into archiveable, mineable texts.

The Early English Books Online dataset (EEBO) consists of a collection of approximately 15 million digital page images of texts published between 1473 and 1700, and these page images are practically impenetrable to OCR engines. Moreover, metadata for such early texts is notoriously unreliable: according to David Foxon, title pages don't only lie, they sometimes joke, naming the printer typically used by a rival author as a way of implicating that author in the text's composition, or naming a bookseller in an area of London such as the theater district, for satirical purposes. Not only the binding of books, but the re-use of previously printed materials in "new" books makes it very difficult to know what is actually proffered by any title--what editions of other works might be included, unacknowledged in the metadata. A consortium of libraries called the Text Creation Partnership (TCP) has decided to key in, type by hand, one instance of each title in the collection, but obviously, in this context, "the same title" rarely indicates what "buried treasures" lie beneath its mark (Jackson). Moreover, it is even the case that individual witnesses of the same edition vary because of stop-press additions and corrections, changes made during the run of a single printing of one edition.

Gibbs muses that "even once we have more reliable OCR technology, it would be nice to have an infrastructure to allow the manuscripts to be viewed together and improved by user expertise" and expresses hope for a transcription editor with an unobtrusive, functional, and intuitive interface, that allows text to be easily (re)configured while displaying variations

between versions. Cobre (COmparative Book REader), a suite of image viewers and tools developed to facilitate detailed interaction with the collection of 16th Century New World imprints in the Proyecto losPrimeros Libros de las Américas: Impresos mexicanos y peruanos del Siglo XVI en la bibliotecas del mundo meets those needs.^[5] Cobre ingests content from an OAI/PMH enabled digital repository. To populate a Cobre instance with texts for eMOP triage, we first structure the page images and their associated OCR transcriptions in the DSpace Simple Archive Format for ingestion into a DSpace repository, from which they can be imported into Cobre. Intrinsic to Cobre's functionality is a Detailed View that not only places page images in context, via a filmstrip metaphor, but provides multiple zoom levels and the ability to drag the page in the viewer pane (Liles et al). Cobre's Comparison View likewise uses a filmstrip view of two or more books together (Liles et al). These filmstrips can be locked, keeping them aligned when any one filmstrip is moved and when a thumbnail in the filmstrip is clicked, a side-by-side view of all the pages appears (Liles et al) in a Quick Comparative View.

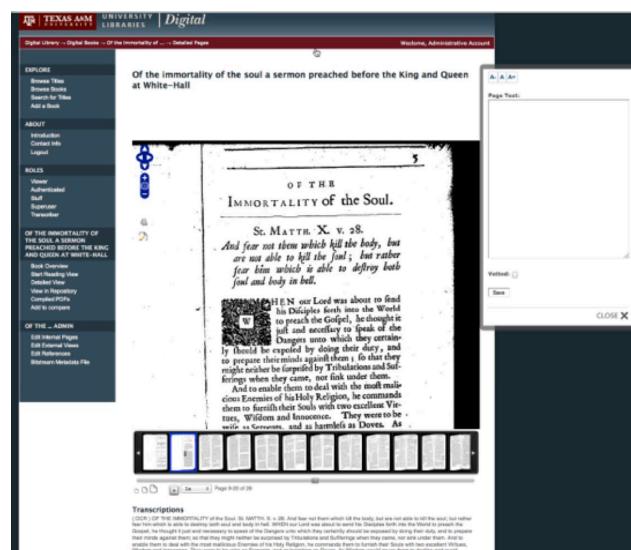


Fig. 1: Cobre's Detailed View with imported OCR and pane for text correction

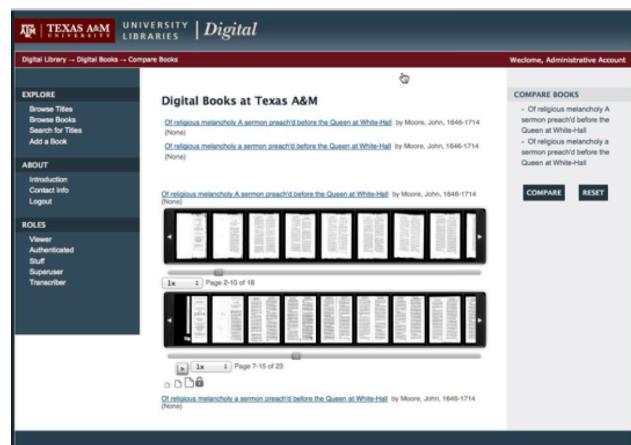


Fig. 2: Cobre's Comparative View of multiples exemplars

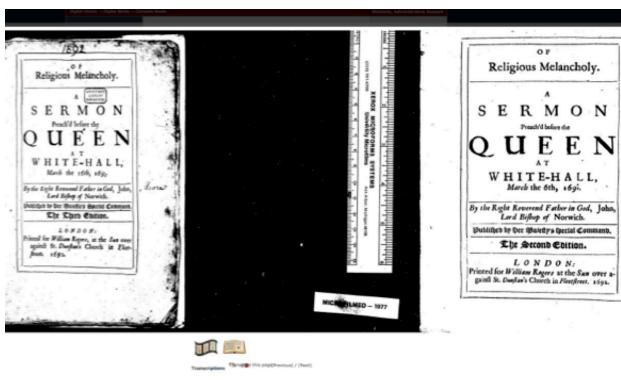


Fig. 3: Cobre's Quick Comparative View of page images and OCR output side-by-side.

Though the Cobre tool was built for the purposes of transcription and thus is technically a crowd-sourced transcription tool like the Bentham wiki and the other tools listed by Melissa Terras,[6] it resembles Ben Brumfield's FromThePage, also mentioned by Terras, in being half transcription tool and half social editor of the sort described by Ray Siemens et. al. Like Bentham, Siemens's own Devonshire ms. uses Wikimedia so that editing can be discussed as well as implemented. Cobre too allows for transcription, annotation, and—a key attraction for experts—editing and adding information to page images and to the metadata.

We performed user studies on the tool, bringing in book history experts James Raven and Robert D. Hume to test the tool, and we videotaped their eye movements while recording their comments. There are many things that they did not find intuitive which we fixed on our last development sprint. We will be performing another set of user studies when we teach approximately 50 people to use Cobre at a pre-conference workshop at the American Society for Eighteenth-Century Studies, to be held in Williamsburg, VA, 19 March 2014.

Transcribe Bentham and the ANDP have been very successful at recruiting experts and citizen scholars (Causer, et. al., and Holley). While the ANDP required a very light form of attention, transcription, and Transcribe Bentham slightly more – users were asked to encode as well – we will be asking users of Cobre to transcribe and compare transcriptions to multiple pages of various editions. Can a network of “authorized” users who will be carefully comparing pages of editions be generated of sufficient strength? Can there be networked—and so massive—close-reading? We will report on whether and how it is possible to generate distributed forms of attention that are required for careful digitization *en masse*.

[1] Kermode's phrase encapsulates the scholarly output over the last two centuries that have produced close readings of the meanings and forms of canonical works of art and literature

[2] Siemens et. al. describe “the social edition,” a procedure that was debated at a recent SSHRC-sponsored conference called “Social, Digital, Scholarly Editing,” hosted by Peter Robinson at the University of Saskatchewan (ocs.usask.ca/conf/index.php/sdse/sdse13). Gibbs calls for a “community transcription tool [that] will reduce significantly the barrier to entry and encourage mark-up of texts,” such as: the CWRC (www.dh2012.uni-hamburg.de/conference/programme/abstracts/cwrc-writer-an-in-browser-xml-editor), FromThePage (beta.fromthepage.com) and “Textual Communities” (www.textualcommunities.usask.ca)

[3] Craig Chappel's view that game-modding is in decline may portend a trend against participation, but that view is arguable.

[4] emop.tamu.edu

[5] primeroslibros.org, libros.library.tamu.edu

[6] melissaterras.blogspot.com/2010/03/crowdsourcing-manuscript-material.html : Scratch, Remote Writer, and the tools hosted by the Australian National Digitisation Program (ANDP) and BYU Historica Journals.

References

- Causer, T., J. Tonra, and V. Wallace** (2012). *Transcription Maximized; Expense Minimized?: Crowdsourcing and editing The Collected Works of Jeremy Bentham*. Literary and Linguistic Computing 27.2, pp. 119-137.
- Causer, T., and V. Wallace** (2012). *Building a Volunteer Community: Results and Findings from Transcribe Bentham*. Digital Humanities Quarterly 6.1. www.digitalhumanities.org/dhq/vol/6/2/000125/000125.html . Accessed 8 January 2014
- Chapple, Craig** (2013). *An FPS Insurgency – Breaking into a Crowded Genre*. 13 September 2013. www.develop-online.net/interview/an-fps-insurgency-breaking-into-a-crowded-genre/0117719 Accessed 31 October 2013.
- Dahlström, Mats.** (2009). *The Compleat Edition*. In Text Editing, Print and the Digital World. Eds. Marilyn and Kathryn Sutherland. Farnham, MA: Ashgate. 27-44.
- Foxon, David, with James McLaverty.** (1991). *Pope and the Early Eighteenth-Century Book Trade*. Oxford: Clarendon Press.
- Gibbs, Frederick W** (2011). *New Textual Traditions from Community Transcription*. Digital Medievalist 7. www.digitalmedievalist.org/journal/7/gibbs
- Gooding, Paul, and Melissa Terras, Claire Warwick** (2013). *The Myth of the New: Mass Digitization, Distant Reading, and the Future of the Book*. Literary and Linguistic Computing 28.4: 629-39.
- Guillory, John** (2010). *Close Reading: Prologue and Epilogue*, ADE Bulletin 149: 8-14.
- Hayles, N. Katherine** (2007). *Hyper and Deep Attention: The Generational Divide in Cognitive Models*, Profession 2007: 187-199.
- Holley, Rose.** (2009) *How Good Can It Get: Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs*. D-Lib Magazine 1.3/4.
- . *Many Hands Make Light Work*. March 2009. National Library of Australia. ISBN 978-0-642-27694-0
- Jackson, Millie.** (2008) *Using Metadata to Discover the Buried Treasure in Google Book Search*. Journal of Library Administration 47.1/2: 165-73.
- Kermode, Frank** (1985). *Forms of Attention*. Chicago: University of Chicago Press.
- Liles, B., Creel, J., Maslov, A., Nuernberg, S., duPlessis, A., Mercer, H., McFarland, M. and Leggett, J.** (2012). *Cobre: A Comparative Book Reader for Los Primeros Libros*. Proceedings of the 45th Hawaii International Conference on System Sciences (HICSS45), Maui, HI, USA. January 4-7, pp. 1707-17. dx.doi.org/10.1109/2fHICSS.2012.155
- Moretti, Franco.** *Conjectures on World Literature*, New Left Review 1: 54-68. (2000)
- . *Graphs, Maps, and Trees: Abstract Models for a Literary History*. New York: Verso, 2005.
- . *More Conjectures*, New Left Review 20 (2003): 73-81.
- . *Style, Inc. Reflections 'Relatively Blunt,'* 172-174.
- Robinson, Peter.** *Textual Communities*. www.textualcommunities.usask.ca
- Siemens, Ray, and Meagan Timney, Cara Leitch, Corina Koolen, Alex Garnet.** *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 (2012): 445-461.
- Terras, Melissa.** (2010) *Crowdsourcing Manuscript Material*. 2 March 2010. Adventures in Digital Humanities Blog. melissaterras.blogspot.com/2010/03/crowdsourcing-manuscript-material.html . Accessed 8 January 2014
- . *Crowdsourcing or crowdsifting? Results and experiences from Transcribe Bentham*. Paper given at Social, Digital, Scholarly Editing Conference, University of Saskatchewan, Saskatoon, Saskatchewan, Canada. 11 July 2013.
- Trumpener, Katie.** (2009). *Critical Response I: Paratext and Genre System*, Critical Inquiry 3.1: 159-71.

The Problem of Time and Space: The Difficulties in Visualising Spatiotemporal Change in Historical Data

Ó Murchú, Tomás
tomas808@gmail.com
Trinity College Dublin

Lawless, Séamus
seamus.lawless@scss.tcd.ie
Trinity College Dublin

The Problem of Time and Space: The Difficulties in Visualising Spatiotemporal Change in Historical Data

Proposal for Digital Humanities 2014 Conference 1 November 2013

Tomás Ó Murchú, MPhil Digital Humanities, Trinity College Dublin, Ireland
Professor Séamus Lawless, Trinity College Dublin, Ireland

Visualisations take advantage of the fact that the human eye has the ability to identify patterns and structures in images that computers are yet to match. Visualisations do this by exploiting features of the human cognitive processing system¹. While much of our communication is done through words, we are connected to our environment primarily through vision. This has resulted in our visual perception having evolved to actively seek meaningful patterns in what we see². Using a digital visualisation system in combination with flexible human cognitive capabilities, such as pattern finding, is far more powerful than an unaided human cognitive process³. Visualising historical data in relation to the information's geographic and temporal attributes can help uncover hidden links and relationships within the data. However traditional spatiotemporal methods for visualising change are often insufficient for providing a spatial and temporal framework within which historical data can be explored. Historians (especially since they normally do not possess the required skillset themselves) have had to live with, or at best modify, existing tools from other disciplines. Because of this they have been channelled down spatiotemporal visualisation routes that are frequently a poor fit for their research.

This paper takes murder information that has been extracted from the 1641 Depositions (testimonies documenting the experiences of witnesses of the 1641 Irish rebellion)⁴ as a case study in creating a spatiotemporal visualisation using historical data. An existing online example that uses the data in an interface with Google Maps is taken as a starting point (downsurvey.tcd.ie/1641-depositions.php).

Fig. 1: Murder information from the 1641 Depositions

Using the same data from the 1641 Depositions, a spatiotemporal visualisation is created to illustrate the difficulties in using historical information for this process.

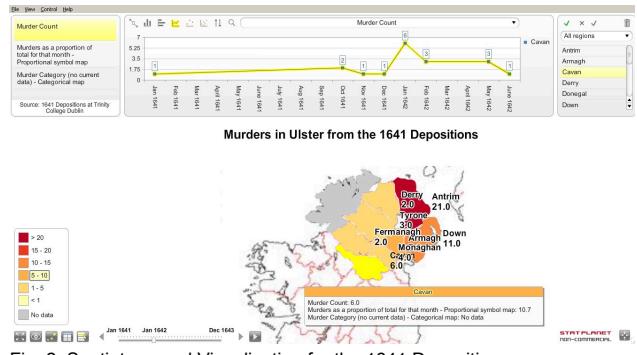


Fig. 2: Spatiotemporal Visualisation for the 1641 Depositions

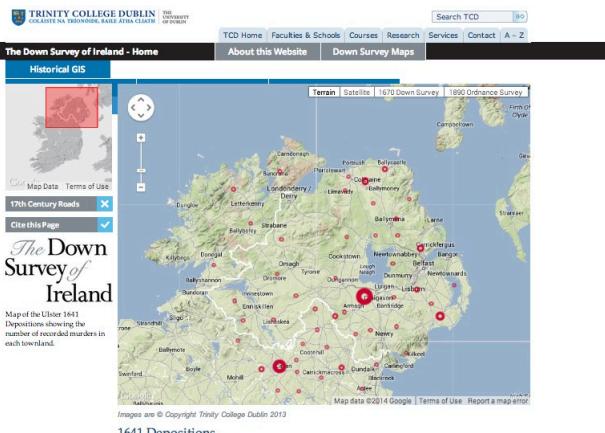
This paper will investigate the problems associated with traditional spatiotemporal visualisations of historical data. It will examine our own comprehension of time and space and how understanding their vagueness, ambiguities and uncertainties are important when it comes to visualising and modelling changes in their components.

Most historical data has a spatiotemporal element to it. This could involve the movement of people over a temporal period within a specific area or the changing area of a political entity over time. As these changes are normally recorded in written texts, historical spatiotemporal data can often exhibit a high degree of uncertainty. Texts are descriptive and by their very nature are vague and open to various interpretations. This qualitative nature of historical documents means that creating a spatiotemporal visualisation involves overcoming obstacles where descriptions of spatial areas or temporal periods are often vague or uncertain. The problem of ambiguity and uncertainty in data is an issue that visualisations do not deal with particularly well⁵. When analysing spatial and temporal data historians need to be conscious of the uncertainties present within the data. Some of these ambiguities may not be immediately apparent and will require detailed analysis to identify.

In historical data there are several reasons why uncertainty occurs. Historians often use data that was intended for a different end use than the analysis that they are trying to accomplish. Trying to convert the data into something usable for a spatiotemporal visualisation inevitably leads to a degree of data compromise. The entities, times and spatial areas mentioned in texts were often meant purely as descriptions related to a specific event so are frequently only vaguely defined. Sometimes the historical sources are transcriptions or translations of lost documents and uncertainties may have occurred due to the transcription or translation process. Similarly, as sources may be transcriptions of oral depositions, cultural, educational and linguistic differences between the transcriber and deponent may cause misinterpretations.

How spatiotemporal data is modelled for spatiotemporal visualisations is an important factor when dealing with historical information. Data models are the conceptual core of an information system. Models should be designed to deal with uncertainties and to create meaningful visualisations of changes over time. Modelling the data includes defining data object types, relationships, operations and rules to maintain database integrity⁶. The data model needs to support the processes that the system will be required to carry out. When modelling spatiotemporal data, a further consideration is the fact that territorial structures and units change over time. The issue of situating (and thus visualising) the data within the correct spatial area at the correct moment in time needs to be effectively managed by the data model.

For historians trying to visualise spatiotemporal information, a data model needs to be able cope with uncertainty and changes over time. Additionally trying to create links between vague concepts in a visualisation is fraught with uncertainty. The



data structure supporting the visualisation can be too rigid to show anything other than a few select relationships. One method to try and reduce and manage uncertainty in modelling historical data is to use what is known as fuzzy logic and fuzzy set theory. The aim of fuzzy logic is to provide a means to cope with ambiguous entities without losing a record of the ambiguity⁷. This is achieved by using assessment rules that the researcher pre-defines and makes transparent so that they can be easily understood and evaluated by other researchers. Fuzzy logic and fuzzy set theory can be used for linguistic variables thus making it suitable for the modelling of textual sources. If used properly it can handle vague and uncertain linguistic labels such as 'slightly', 'close to' or 'very' (as in 'is very old'). These linguistic variables are present everywhere in written and spoken communication but computers have difficulty in recognizing their correct application.

Modelling qualitative spatiotemporal information provides many challenges. The computational nature of traditional visualisation systems mean that information needs to be categorised into set groups. Historians often resist such demands as they feel that some vital characteristics or attribute inherent in the language describing it will be lost when they do not fit precisely a particular category⁸. The context in which an entity is described can be as important as the attributes of the entity itself. A death in a source text may be described as a murder by one witness, an accident by another or an act of self-defence by another. Trying to categorise it or by assigning it neutral label such as 'unnatural death' robs the entity of all meaning in a historical context. Extracting data from historical textual sources by computational means is likely to miss out on much key contextual data. Domain experts understand that there is meaning present in descriptions that can defy conventional numerical and computational approaches.

Overcoming uncertainties and vagueness in the Deposition texts proved challenging for existing spatiotemporal visualisation tools. In a text such as the 1641 Depositions, the multitude of ways dates are represented causes huge difficulties in linking particular events to the date they occurred. Phrases such as 'at the beginning of lent of that year' and 'two months hence' abound. Representing all these different time periods on the same visualisation is very difficult. There may also be overlaps between the periods of time with one text identifying a particular event on a particular day while another text in the canon identifying the exact same event but it occurring somewhere in the period of a month. Similarly, identifying places in the Depositions is also problematic with many places referred to in uncertain terms. General terms such as 'near', 'close to' or 'in the region of' are used extensively in the texts. Anglicizations of Gaelic place names in the Depositions are inconsistent and often do not correlate to modern spellings.

Finally, future work on how Linked Data and the Semantic Web may have the ability to help historians to overcome spatiotemporal visualisation limitations is considered. The paper concludes that new tools and data models are required to effectively visualise spatiotemporal historical information.

References

1. Keller, Tanja, and Sigmar-Olaf Tergan (2005). *Visualizing knowledge and information: An introduction*. Knowledge and information visualisation p. 5.
2. Nöllenburg, Martin. (2007) *Geographic visualization*. Human-Centered Visualization Environments p. 257.
3. Keller, Tanja, and Sigmar-Olaf Tergan. (2005) *Visualizing knowledge and information: An introduction*. Knowledge and information visualization p. 5.
4. www.1641.tcd.ie
5. Amar, Robert A., and John T. Stasko (2005). *Knowledge precepts for design and evaluation of information visualizations*. Visualization and Computer Graphics, IEEE Transactions p. 433.
6. Yuan, May. (1996) *Temporal GIS and spatio-temporal modelling*. Proceedings of Third International Conference Workshop on Integrating GIS and Environment Modelling p.1.
7. Owens, J.B. and Coppola, Emory J. (2012) *Fuzzy Set Theory (or Fuzzy Logic) to Represent the Messy Data of Complex Human (and other) Systems*. White Paper Presented to the US National Science Foundation p.2.
8. Owens, J.B. and Coppola, Emory J. (2012) *Fuzzy Set Theory (or Fuzzy Logic) to Represent the Messy Data of Complex Human (and other) Systems*. White Paper Presented to the US National Science Foundation p.2.

Posters

Mapping French Press to the Digital Age

Abi Haidar, Alaa

alaa.abi-haidar@lip6.fr

University of Pierre and Marie Curie

Ganascia, Jean-Gabriel

jean-gabriel.ganascia@lip6.fr

University of Pierre and Marie Curie

Abstract

The unsupervised use of dictionary-lookup is known to enhance NER, however dictionaries have limitations for being finite and ambiguous. On the other hand, supervised NER such as Stanford's NER Classifier that we tested here is known to perform very well but only with the availability of huge amounts of manually annotated training data that is very costly, time consuming and sometimes inaccurate due to inter-annotator inconsistencies. Therefore, we develop and discuss an original unsupervised approach for Named Entity Recognition (NER) and Disambiguation (UNERD) using a French knowledge-base and a statistical contextual disambiguation technique that slightly outperformed Stanford's NER Classifier (when trained on a small portion of manually annotated data) and Aleda's dictionary lookup. Furthermore, we devise a solution to identify and highlight named entity tagging on the original scanned images thus preserving the newspaper's layout and feel.

1. Introduction

Huge amounts of printed manuscripts from old French journals (from the 19th and 20th century) have been recently digitized and published by the National French Library, la Bibliothèque Nationale Francaise (BnF). However, the massive amounts of produced textual data are highly unstructured and hard to index or search, needless to mention the digitization errors resulting from ill-preserved or damaged manuscripts and imperfect Optical Character Recognition (OCR) techniques.

Named Entity Recognition (NER) is a task of information extraction 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. Furthermore, advanced NER and disambiguation techniques are capable of dealing with noise resulting from digitization errors.

Several supervised learning techniques, such as Stanford's Conditional Random Field (CRF) Classifier¹, have been developed to address the question of NER very successfully, however they require large manually-annotated corpora of text, which is very expensive to obtain and maintain. On the other hand, with the abundance of publicly accessible knowledge bases and dictionaries, such as Freebase, geonames and Aleda², unsupervised methods have become popular especially since they require no pre-annotated training data. However, several challenges have arisen from the choice of appropriate knowledge bases, resolving homonymous ambiguity, detecting entity boundaries, and identifying less common name entities. Moreover, most studies have been dedicated to English text while text in other languages, that has recently been digitized and published at astounding rates, has received little or no attention.



Source gallica.bnf.fr / Bibliothèque nationale de France

Fig. 1: Original scanned image from "Le Petit Parisien"

Here, we discuss our model of Unsupervised Named Entity Recognition and Disambiguation (UNERD), and validate it on digitized French text from the 19th century. We claim that NER using Knowledge based disambiguation is especially relevant for minimally annotated texts that are expensive to annotate in several languages and domains, which is often the case in Digital Humanities. More details about our algorithm and its performance are detailed in³. We also discuss a solution that uses the XML digitized data in order to preserve the location of entities and highlight them on the originally scanned journal image as well as in the OCRised text.

2. Data

Here we discuss the data at hand, its format, encoding and necessary text processing. Next we discuss a portion of annotated data that is used for validation. Finally, we propose a solution that allows tagging the originally scanned journal image.

2.1. Corpus

The corpus consists of recently digitized and published old unlabelled French journals. More specifically, the data consists of a subset of "Le Petit Parisien" journal supplied by Bibliothèque nationale de France (BnF). It was originally published between 1863 and 1944 with a total of 29616 issues. The corpus we are using comprises 260 issues with a total of 1098 pages of natural French text. The pages were recently digitized and OCRised and encoded in the ALTO format, which is an open XML standard for representing OCR text. Both scanned pages and text formats of "Le Petit Parisien" are accessible at the BnF website via their digital portal, Gallica: gallica.bnf.fr/ark:/12148/cb34419111x/date.langFR

2.2. Format

The ALTO file has 3 major sections:

- <Description> contains the metadata about the ALTO file.
- <Styles> specify the text and paragraph styles.
- <Layout> describes the main content subdivided into <page> elements.

Each page of the content is further divided into margins and printspace, each of which containing lines, images or textblocks. The ALTO format provides OCR confidence for each word. The XML files are encoded using the iso-8859-1 encoding. Handling this dataset is challenging due to the old ill-preserved manuscripts resulting in many OCR errors. However, we exclude text blocks with low OCR confidence (lower than 0.85).

2.3. Validation Data

The corpus of French journals is very expensive and time consuming to annotate. Therefore we had experts annotate only a small portion of 4171 words of which the numbers annotated entities are 75 Person, 78 Location, and 22 Organisation entities.

2.4. Text Processing

Text preprocessing is essential for decoding and analysing the data and the first step is to prepare OCR text for entity extraction such as converting the XML iso-8859-1 encoded corpus into standard UTF-8 encoded text, excluding text-blocks with confidence less than 85% or containing words with length more than 15 characters, applying Part of Speech Tagging (using TreeTagger) and removing French stop-words.

2.5. Representation

Here, we propose a solution that enables the ALTO XML format to keep track of the location of the words in order to highlight the entities not only for the OCRised text but also for on the original scanned image thus maintaining the layout, font types and feel or reading a newspaper.

The ALTO XML format preserves the location of all OCRised words using the tags HPOS and VPOS as illustrated in the following code for the text "Toussaint, humide et pluvieux":

```
<Textline ID="PAG_1_ST000000" STYLEREF="TXT_1" HPOS="60" VPOS="331" HEIGHT="22" WIDTH="602">
<String ID="PAG_1_ST000000" STYLEREF="TXT_1" HPOS="60" VPOS="331" HEIGHT="22" WIDTH="602" NC="0.98" CONTENT="Toussaint,"/>
<String ID="PAG_1_ST000000" HPOS="149" VPOS="332" WIDTH="23"/>
<String ID="PAG_1_ST000000" HPOS="149" VPOS="332" HEIGHT="20" WIDTH="103" NC="0.99" CONTENT="humide,"/>
<String ID="PAG_1_ST000000" HPOS="149" VPOS="332" WIDTH="21"/>
<String ID="PAG_1_ST000000" STYLEREF="TXT_1" HPOS="566" VPOS="332" HEIGHT="20" NC="0.99" CONTENT="et,"/>
<String ID="PAG_1_ST000000" HPOS="566" VPOS="331" HEIGHT="24" NC="0.99" CONTENT="pluvieux,"/>
```

Fig. 2: Sample ALTO XML code representing 4 words and their positions on the scanned image



Fig. 3: Sample original scanned image from "Le Petit Parisien" highlighting entities corresponding to the following categories: Location (blue), Organization (green) and Person (red) names.

3. Method

Our algorithm first uses a French knowledge-base, namely Aleda, to find contextual collocations (or word co-occurrences) for each entity class (Person, Location or Organisation) within a window of size n. Next, we compare the context of the

candidate entity with the classes' contextual cues that eventually suggest a class name for each contextual word around the entity-mention. The comparison may or may not rely on the position of the three neighboring words that make its context based on the following disambiguation techniques.

- The single significant cue selects the class of the contextual cue with the highest TF-IDF score.
- The bag of words selects the class of the contextual cue with the highest TF-IDF ignoring the relative positions of the contextual cues.
- The majority rule selects the class that has the majority of the votes by contextual cues or at least two out of three votes.

Finally, our algorithm uses the majority rule to select a class based on the results from the previous techniques i.e. if at least two disambiguation techniques agree on a class then this class is chosen. Furthermore, the entity-boundary detection uses a Parts of Speech (PoS) tagger that identifies noun phrase boundaries.

3. Results

We tested our UNERD algorithm on a small portion of data that we had annotated by experts and we compared it using k-fold cross-validation to mere dictionary look-up, namely Aleda, and to Stanford's NER Conditional Random Field (CRF) Classifier. The preliminary results reported using F-score are clearly in favor of our unsupervised method as shown in table 1.

F-score	Aleda Look-up	Stanford NER	UNERD
LOC	0.33	0.54	0.77
PER	0.73	0.75	0.83
ORG	0.20	0.44	0.46
AVERAGE	0.42	0.58	0.69

For more details about K-fold cross-validation, F-score please refer to ⁴.

Conclusion

Here, discussed an original unsupervised named entity recognition and disambiguation approach which outperforms mere dictionary lookup and supervised learning on small portions of annotated data. Arguably, Stanford's supervised NER can outperform our method if trained on a larger set of manually annotated data. However, that may be true with the availability of huge amounts of annotated data that are very expensive and time consuming to produce. Our NER and disambiguation technique is statistical and is supposed to work on various languages and domains using no annotated data. We also discussed an approach for keeping track and making use of the ALTO XML format in order to highlight entities on the original scanned image.

References

1. Finkel, Jenny Rose, Trond Grenager, and Christopher Manning (2005). *Incorporating non-local information into information extraction systems by gibbs sampling*. Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics.
2. Benoit Sagot, Rosa Stern, et al. (2012). *Aleda, a free large-scale entity database for french*. Proceedings of LREC 2012
3. Mosalleem, Yusra, Abi-Haidar, Alaa and Ganascia Jean-Gabriel (Submitted). *Unsupervised Named Entity Recognition and Disambiguation: An Application to Old French Journals*.
4. Feldman, Ronen, and James Sanger (2007). *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge University Press

A Digital Metaphor Map for English

Anderson, Wendy

University of Glasgow, United Kingdom

Aitken, Brian

University of Glasgow, United Kingdom

Hamilton, Rachael

University of Glasgow, United Kingdom

In this poster, we will outline a new digital resource, the 'Metaphor Map' of English, which is opening up empirical research into metaphor on a scale never before possible. We will also set out the new methodology which underpins the resource, and outline its significance for research and teaching in metaphor studies and in Digital Humanities.

Metaphor is pervasive in language and is a focus of research in many disciplines including linguistics (e.g. semantics, discourse analysis, historical linguistics), history of ideas, psychology, literature and cognitive science. Linguistic interest in recent decades, following the groundbreaking work of Lakoff and Johnson (1980), has focused on metaphors in everyday language, such as the systematic connection between heat in the material world and abstract concepts of anger or emotion, shown in expressions such as 'fuming' or 'inflamed'. However, the lack of a comprehensive data source has made it difficult to obtain an overview of this phenomenon for the history of English.

The recent completion, after 40 years of research at the University of Glasgow, of the *Historical Thesaurus* (HT) database, published as the *Historical Thesaurus of the Oxford English Dictionary* (Kay, Roberts, Samuels and Wotherspoon 2009), has now made it possible to attain this overview. A unique contribution to language scholarship, the HT is organised in hierarchical semantic categories, each containing lists of words used to express given concepts at particular points in time. The HT charts "the semantic development of the huge and varied vocabulary of English" (Kay et al. 2009: ix), and can provide a snapshot of lexical usage at any given time: scholars may thus compare historical links between categories from a new perspective, gaining fresh insights into how the language has developed. The underlying data sources are the 2nd edition of the *Oxford English Dictionary* for the period from 1150 to the present, and *A Thesaurus of Old English* (Roberts and Kay 2000) for 700-1150. The HT contains 793,742 meanings, hierarchically organised into 225,131 conceptual categories. Its primary division into three main sections, the external, mental and social worlds, lends itself to identifying metaphorical transfer from concrete to abstract.

The HT forms the source data for the 'Mapping Metaphor with the *Historical Thesaurus*' project, funded by a three-year Arts and Humanities Research Council grant (Jan 2012 – Dec 2014, PI - Anderson). The principal aim of Mapping Metaphor is to undertake an empirical investigation of the foundations and nature of metaphor in English. As a first step, we ran a series of automatic routines to identify lexical overlap between semantic categories, and created a database providing a comprehensive mapping of recurrent words. Computational methods for determining relative significance of lexical overlap were explored, but were found to be of limited value and did not obviate the need for detailed manual analysis of the complete data set to locate and annotate metaphorical connections between semantic categories. Subsequent qualitative analysis of these data has provided a platform from which to assess the extent and nature of metaphorical links in the history of English.

An interactive digital Metaphor Map will be made freely available online (estimated launch date, late summer 2014), and will be demonstrated at the conference. The interface to the Metaphor Map resource was created using the D3 (Data Driven Documents) Javascript visualisation library, and the poster will present a technical overview of the visualisations, the reasoning behind the interface that was developed and a brief discussion of some of the alternative approaches that were investigated during the development process. The Map is designed specifically for researchers and members of the public to browse the identified links, cross-refer between

semantic categories linked metaphorically, and drill down to the underlying lexical data. Users can locate a category of interest either through a keyword search or by browsing the highest level of the Metaphor Map. For example, the category "Colours" is found under "Matter" and the Metaphor Map reveals a number of metaphorical connections between "Colours" and other categories. Categories which have a strong, systematic metaphorical link with "Colours" include: "Age", "Sexual relations", "Fear", "Virtue" and "Moral evil". Users can then navigate to a page which will give a list of sample lexemes of overlap between two categories selected. Here users can view a snapshot of the underlying data, such as how the colours *black*, *purple* and *scarlet* appear in "Moral evil" with the sense of 'wicked'. Categories which have a weaker metaphorical connection are also indicated on the Metaphor Map and a list of semantically similar categories is given, which will help to guide users towards other categories of interest based on their current category; for example, when searching for "Colours", users will also be directed towards: "Chemistry", "Light" and "Variegation".

The poster will also explore the significance of the project and the Metaphor Map web resource, both for Digital Humanities and for the field of metaphor studies. In a DH context, Mapping Metaphor is significant because it represents a project building on the successes of a previous project, the *Historical Thesaurus*, which had been born in the 1960s before the widespread use of computers in the humanities, but kept pace with the burgeoning digital landscape through the 1980s and 1990s.

Metaphor is one of the main agents in the development of polysemy and semantic change. However, previous data sets have not been able to show the role of metaphor on a large scale. Semantic change driven by metaphor is traditionally assumed to proceed from concrete to abstract domains. For example, in the conceptual metaphor SOCIAL ORGANISATIONS ARE PLANTS, the source domain, 'plants', is more concrete than the target, 'social organisations'; the abstract sense 'branch' (of a bank) developed from the concrete sense 'branch' (of a tree). The Mapping Metaphor project is allowing us to identify large-scale patterning, and giving us a context within which to explore possible counter-examples. Allan (2008: 186), in a study using HT data, notes that "While cases of concrete to abstract metaphorical mapping are undoubtedly more common, an increasing number of semantic shifts in the opposite direction have been documented". She also draws attention to cases where source and target concepts appear not to have been separate historically: the inclusion in the HT of dates/spans of attestation for all senses allows for full investigation of this further complication. Likewise, considering the data synchronically, the Metaphor Map enables exploration of the HT's information on start-dates of lexical usages to obtain a snapshot of lexis and metaphor at key points in the history of English.

Kövecses (2010:27) has remarked that "More precise and more reliable ways of finding the most common source and target domains are needed". Previous research, while empirically-based (e.g. exploiting dictionaries and synchronic thesauruses), has not been comprehensive, so it has not been possible to explore fully statements like Sweetser's (1990:18): "in some cases there is a deep cognitive predisposition to draw from certain particular concrete domains in deriving vocabulary for a given abstract domain". HT data and the methods developed in the Mapping Metaphor project are making possible a near-comprehensive identification of such domains.

The resource also opens up a number of further areas of investigation for scholars within metaphor studies. For example, the use of the vast HT, with its hierarchical, conceptual organisation, presents a new perspective on the current debate concerning the validity of the notion of domain itself (e.g. Panther 2006, Geeraerts 2010). In reconsidering domains, our methodology opens up questions concerning the relationship between metaphor and metonymy, traditionally defined as extra- and inter-domain mapping respectively. HT data can also be used to explore the relationships which metaphor holds with part of speech and register.

The core of the project deals with metaphor as it is encoded in the language system and evidenced in the HT: however, the

approach taken here is informed by and complements work in the two other current main avenues of research, metaphor in text (e.g. Pragglejaz Group 2007, Deignan 2006) and metaphor in human cognition (e.g. Lakoff 1987, Lakoff and Johnson 1980, 1999). We anticipate that the Metaphor Map resource will also prove to be of value to researchers and teachers in these areas.

References

- Allan, K.** (2008). *Metaphor and Metonymy: A Diachronic Approach*. Oxford: Wiley- Blackwell
- Deignan, A.** (2006). 'The grammar of linguistic metaphors'. Stefanowitsch & Gries (eds). *Corpus-based Approaches to Metaphor and Metonymy*, 106-122. Berlin/New York: Mouton
- Geeraerts, D.** (2010). *Theories of Lexical Semantics*. Oxford: OUP
- Kay, C., J. Roberts, M. Samuels and I. Wotherspoon** (eds) (2009). *Historical Thesaurus of the Oxford English Dictionary*. Oxford: Oxford University Press
- Kövecses, Z.** (2010). *Metaphor: A Practical Introduction*. 2nd edition. Oxford: OUP
- Lakoff, G.** (1987). *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*. Chicago: University of Chicago Press
- Lakoff, G. and Johnson, M.** (1980). *Metaphors We Live By*. Chicago: University of Chicago Press
- Lakoff, G. and Johnson, M.** (1999). *Philosophy in the Flesh*. New York: Basic Books Oxford English Dictionary online, <<http://www.oed.com>>
- Panther, K.-U.** (2006). 'Metonymy as a usage event'. Kristiansen, Achard, Dirven, Ruiz de Mendoza Ibáñez (eds). *Cognitive Linguistics: Current Applications and Future Perspectives*, 147-186. Berlin/New York: Mouton
- Pragglejaz Group** (2007). 'MIP: A method for identifying metaphorically used words in discourse'. *Metaphor and Symbol* 22(1):1-39
- Roberts, J. and Kay, C. with Grundy, L.** (2000). *A Thesaurus of Old English*, 2 vols. 2 Amsterdam: Rodopi
- Sweetser, E.** (1990). *From Etymology to Pragmatics*. Cambridge: CUP edn. Oxford: OUP

Quantifying "The Thing Not Named": A Computational Analysis of Willa Cather's "Unfurnished" Writing Style(s)

Ankenbrand, Rebecca

rankenbrand@gmail.com

University of Nebraska, United States of America

Bernardini, Caterina

aterina.bernardini@gmail.com

University of Nebraska, United States of America

Brotnov Eckstrom, Mikal

mikalbeckstrom@icloud.com

University of Nebraska, United States of America

Kinnaman, Alex

aokinnaman@gmail.com

University of Nebraska, United States of America

Tedrow, Kimberly Ann

ktedrow2@unl.edu

University of Nebraska, United States of America

Overview:

Willa Cather enunciated the theoretical foundations of her writing style in her 1922 *ars poetica* essay "The Novel *Démeublé*" (Transl. "The Novel Unfurnished"). In this work,

Cather calls for a realistic writing style that is rooted in an economy of prose still rich in suggestion and emotions. Cather describes a writing style capable of creating a bare scene where literalness ceases and where readers can detect the presence of what she calls "the thing not named." Such expositions prompt us to ask questions about what it means to *unfurnish* literary piece and to wonder if the writing style Cather described in this seminal essay is present in her own fiction—either throughout her entire fictional corpus, or only after her manifesto.

These questions provide the impetus for our computational study of Cather's fiction, and the results of our analysis provide insight into the question of how we evaluate whether the author's production is congruent with the author's *ars poetica*. Starting from Cather's declarations in "The Novel *Démeublé*," we defined the criteria for a lexical and syntactical analysis of her fiction. Our analysis indicates that Cather's writing style remained consistent across her corpus both before and after her stylistic declaration in "The Novel *Démeublé*," providing new answers to the long-standing scholarly debate as to whether Cather ascribed to the ideas in the essay.

Methodology:

In this work, we use high frequency words and selected parts of speech to explore whether a chronological shift from a more ornate style to a minimalist style can be detected across a career spanning forty-eight years. Our team utilized over sixty examples of Willa Cather's fictional work currently housed at the Willa Cather Archive in XML format. The textual data was composed of two "genre sets." The first set included all of Cather's novelistic fiction and the second, Cather's short story fiction. Using R (R Core Team, 2013), our team parsed, tokenized, and POS tagged all of the XML into word and parts of speech frequency tables which could then be studied chronologically and in isolation.

Using this derivative data, our team specifically studied Cather's use of determiners, adverbs, adjectives, and personal pronouns as a way of measuring Cather's idea of unencumbered prose. A more "unencumbered" prose, we hypothesized, would be less ornate and utilize fewer of these markers. To facilitate this investigation, we calculated the mean for each selected parts of speech and used those means as a basis for further examining outliers within the long form and short form fiction. Among other things, our team compared the part of speech frequencies between Cather's long form fiction (Fig 1) and short form fiction (Fig 2) and then specifically analyzed pronouns (Fig 3), determiners (Fig 4), and adverbs within the respective corpora (Fig 5).

POS	Death	LFF
Personal pronouns	5.56	7.98
Determiners	9.97	8.34
Adverbs	5.44	4.55

Fig. 1: Parts of Speech in Death Comes to the Archbishop with Cather's Long form fiction (LFF)

POS	White	SFF
Personal pronouns	3.97	7.99
Determiners	13.49	8.5
Adverbs	4.1	5.86

Fig. 2: Parts of Speech in Tale of the White Pyramid with Cather's short form fiction (SFF)

Personal pronouns

	White/SFF	Death/LFF
• He	1.05/1.87	2.0/1.81
• She	0	0.29/1.27
• Him	0.48/.61	0.56/0.56
• Her	0	0.4/1.3
• It	1.1/1.15	.88/1.5

Fig. 3: Use of top five personal pronouns in Tale of the White Pyramid against the mean in Cather's short form fiction (SFF) and in Death Comes to the Archbishop versus the mean in Long form fiction (LFF)

Determiners		
	White/SFF	Death/LFF
• A	1.54/2.0356	2.4/2.21
• That	.6/1.24	1.12/1.7

Fig. 4: Use of top two determiners in Tale of the White Pyramid against the mean in Cather's short form fiction (SFF) and in Death Comes to the Archbishop versus the mean in Long form fiction (LFF)

Adverbs		
	White/SFF	Death/LFF
• Very		
• Much		
• Not	.349/.51	0.55/0.45
• Also		
• Too		

Fig. 5: Use of adverbs in Tale of the White Pyramid against the mean in Cather's short form fiction (SFF) and in Death Comes to the Archbishop versus the mean in Long form fiction (LFF)

For our analysis of the high frequency words, we created a minimum mean threshold and examined those words within a minimum relative frequency across the corpus of 0.5. This threshold had the effect of excluding context sensitive words from the analysis. All of this derivative data was merged with metadata from the texts and then explored and organized using the Euclidian metric as a basis for a hierarchical clustering. The clustering allowed us to investigate the similarities between sixty-six samples of Cather's fictional writing. A dendrogram (Fig 6) shows the results of this clustering.

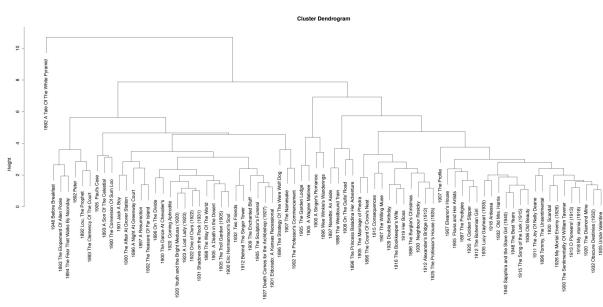


Fig. 6: Word Frequency Clustering of Entire Fictional Corpus

Conclusion:

Our analysis of Cather's use of specific parts of speech and high frequency words indicated no significant change in her usage patterns over time. Our results suggest that the style Cather advocated in "The Novel Démeublé" is the style that she employed throughout her career. Despite a generally stable signal over time, though, two outliers emerged: *Death Comes to the Archbishop* and *A Tale of the White Pyramid*. The former was a distant outlier within a closed set of her novelistic fiction but was found to be consistent with her shorter fiction. The latter was the sole outlier of the entire corpus, which is not surprising given that Cather wrote it as a student.

References

- E. K. Brown (2003) "Homage to Willa Cather," in Willa Cather Critical Assessments, Guy Reynolds, ed. Mountfield: East Sussex. Vol. I
- Maciej Eder "Stylo R Package." sites.google.com/site/computationalstylistics/stylo
- David Hoover (2003), "Another Perspective on Vocabulary Richness," Computers and the Humanities, 37:15'-178.
- Mary Ann O'Farrell (2005), "Words To Do with Things", in Willa Cather and Material Culture, Janis P. Stout, ed., University of Alabama Press: Tuscaloosa.

David Oshinsky (2007), "No Thanks, Mr. Nabokov," New York Times, September 9. www.nytimes.com/2007/09/09/books/review/Oshinsky-t.html?_r=0 .

R Core Team (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, www.R-project.org .

Lionel Trilling (2003), "Willa Cather," in Willa Cather Critical Assessments, Guy Reynolds, ed. Mountfield: East Sussex. Vol. I.

This research began in a class taught by **Matthew Jockers** and has continued under his direction as a project of the Nebraska Literary Lab.

See: Willa Cather Archive. "The Novel Démeublé" cather.unl.edu/nf012.html . First published in the New Republic, 30 (April 12, 1922): 5-6.

Le labo junior « Nhumérisme » (ENS Lyon), observateur et acteur du « cultural empowerment » français

Armand , Cécile

ENS Lyon - Labo junior Nhumérisme, France

Un labo junior est une structure de recherche interdisciplinaire unique créée par l'ENS de Lyon. Il a pour vocation de donner à des doctorants et jeunes chercheurs, issus d'horizons disciplinaires et géographiques variés, une première expérience de la recherche, dans toutes ses dimensions : depuis la mise en œuvre d'un projet scientifique collectif, que dans ses aspects les plus pratiques, logistiques voire « politiques » (gestion et de « management » d'une équipe de recherche, démarches de financement). Crée en avril 2013, le tout jeune labo junior « Nhumérisme », d'abord dédié à des « humanités numériques » mal définies, tend à glisser vers un « humanisme numérique » à construire. Le néologisme « nhumérisme », né de la contraction entre « numérique » et « humanisme », postule que le « numérique », au-delà des outils et d'une dimension purement technique voire technicienne, est un fait de société et de civilisation, un processus qui reconfigure nos héritages, nos savoirs et nos pratiques, nos valeurs, nos rapports aux autres et au monde. Plus qu'un fait, la civilisation numérique nous apparaît comme un projet, une exigence éthique et pratique. A ce titre, notre « labo junior » s'inscrit doublement dans ce phénomène de « *cultural empowerment* » : non seulement en tant qu'observateur, mais aussi comme acteur même de cette « montée en puissance du numérique ».

Nhumérisme, observateur critique du *cultural empowerment*.

Le numérique en est l'objet d'étude même de notre labo junior, dans une perspective critique : loin de le considérer comme allant de soi, nous cherchons dans le cadre de nos divers ateliers et séminaires, à mettre en question sa définition et son existence même, à en retracer l'histoire et en cerner les perspectives, à replacer cette « montée en puissance du numérique » dans le temps long de l'histoire des techniques et de la culture, pour faire la part des permanences et des mutations, et déconstruire le mythe d'une « révolution numérique ». Si cette démarche épistémologique se nourrit d'une culture « humaniste » ancienne, de compétences « traditionnelles » acquises au cours de nos cursus en sciences humaines et sociales, nous tentons aussi de définir et d'acquérir de nouvelles « littératies » pour évoluer dans ce nouvel environnement numérique. Nous pensons qu'il est nécessaire de recomposer nos héritages, de développer des savoirs, savoirs-faire et compétences spécifiques permettant de s'approprier les nouveaux outils et médias, afin de rester maître de la production et la diffusion des connaissances scientifiques. Ainsi, notre première journée d'étude intitulée « la tour de Babel

numérique » (octobre 2013) visait à appréhender la diversité des approches et des « langues » autour du numérique, pour en éprouver l'unité et en proposer une définition plurielle. Par la suite, des conférences plus « thématiques » cerneront des domaines plus ciblés : bandes dessinées numériques (novembre 2013) ; *design* et SHS (décembre 2013), réseaux (février 2014) ; livres, écritures et lectures à l'ère numérique et digitale (avril 2014), datavisualisation (dans le cadre du That Camp Lyon - octobre 2014). Notre premier atelier intitulé « Mapping DH » (novembre 2013) réfléchit aux questions de nomination, à l'élaboration d'un dictionnaire critique ou d'une encyclopédie des humanités numériques, collaboratifs et actualisables, une cartographie, une sociologie des humanités numériques (ses lieux, ses institutions, ses réseaux, ses acteurs), et une histoire des digital humanities, ainsi qu'une bibliothèque virtuelle des grandes références et ressources en humanités numériques.

Nhumérisme, acteur du *cultural empowerment*.

Nhumérisme est un bon observatoire pour les sociologues des sciences qui étudient la fabrique des projets « numériques » dans leur complexité voire leur ambivalence. Sorte de projet « éprouvette », notre labo junior se sert des institutions tout les servant (mais sans leur être asservi...). Bien que labellisé et financé par l'ENS de Lyon, qui a perçu dans l'étiquette « numérique » de notre projet une manière d'accroître sa visibilité et de suivre un effet de « mode », nous disposons toutefois d'une grande liberté dans le choix des thématiques et des actions que nous menons. La souplesse et la jeunesse de la structure autorise des expérimentations originales et des prises de risque qui intègrent la possibilité de l'échec, moteur même de la connaissance scientifique, pourtant condamné par les politiques de recherche actuelles, les logiques et la temporalité mêmes des projets (court-termisme, obligation de résultats). Acteurs et même médiateurs du *cultural empowerment*, nous jouons un rôle d'interface entre des communautés séparées, entre projets et des structures épars. Au sein du monde académique d'abord, nous faisons souvent le lien entre différents métiers (chercheurs, ingénieurs ou techniciens), laboratoires, disciplines, générations (junior/ senior). Nous tentons également d'établir des ponts entre le monde académique, d'un côté, et le monde non académique, de l'autre (artistes et *designers*, entreprises et décideurs politiques, grand public et citoyens). Nous pensons que le numérique peut être une opportunité pour le chercheur de « sortir de sa tour d'ivoire », de travailler et collaborer avec d'autres professions, de réaffirmer son importance sociale et de réveiller l'intérêt du public pour la fabrique de la connaissance scientifique en SHS. Au risque d'un angélisme voire d'un évangélisme numérique, peut-être naïf et moralisateur, mais pleinement assumé, nous pensons les *digital humanists* investis d'une mission critique et pédagogique pour accompagner les citoyens que nous sommes face au déferlement numérique et au danger de l'analphabétisme numérique.

tradition about the Jewish temple in Jerusalem provided the model for these lighting effects as well as for the common use in early medieval churches of windows with jambs and sills that widen on the inside to expand the projection of natural light. Archaeoastronomers have hypothesized that select medieval pictorial programs were coordinated with fenestration to spotlight specific scenes and figures on specific days and at specific hours. We have created a 3D model that visualizes passage of sunlight on any particular day onto and across the walls of the monastery of Saint John in Müstair, Switzerland, the earliest standing church for which such coordination has been proposed. Our model tests and refines the theory of Gion Gieri Coray-Lauer, a Swiss archaeoastronomer.

The history of the region and other notable features of the church make Coray-Lauer's hypothesis highly attractive. First, early medieval churches in the region were commonly aligned with still-standing prehistoric markers delineating astronomical lines or with the rising sun on a patron's feast day. Second, the windows of the church have highly decorated jambs and sills, and they slant at various angles according to the direction that they face. Third, the feast days of the saints highlighted in the three apses of the church cluster around the summer solstice, when persistent pagan practices denounced by local preachers culminated each year.

Although situated in a remote alpine valley, the main church of the Monastery of Saint John preserves the most extensive program of church decoration in the west to survive from the first millennium. The program includes the earliest preserved monumental Last Judgment, east or west, among many precocious pictorial themes. The highly visible wall paintings display great clarity and exemplify the dictum of Pope Gregory the Great that pictures could serve as books for the illiterate.

A greater understanding of the lighting effects deepens our understanding of the multi-sensory experience of medieval church decoration as well as our understanding of a medieval monument of great importance both for the extensiveness of its pictorial program and for the precocity of many images within it.

Scholars have well explored the symbolism of light in Christianity, both in text and image. Writers, both medieval and modern, have also written about the potential for light to move and dazzle the worshipper. Older scholarship on the quality of light in churches was based entirely on observation, but the methods and calculations of archaeoastronomy, which are generally incomprehensible to humanities scholars, have yet to penetrate the mainstream of art historical scholarship.

The visualization of archaeoastronomical data within 3D virtual models enables art historians to judge more easily archaeoastronomical theories and to incorporate them into interpretations of how architects and designers of decorative programs structured and shaped religious experience.

We propose to demonstrate our software application in an interactive poster session.

References

iav.ipfw.edu/iu_home.html

Interoperable Infrastructures for Digital Research: a proposed pathway for enabling transformation

Baker, James

james.baker@bl.uk
British Library

Farquhar, Adam

Adam.Farquhar@bl.uk
British Library

Governments, research organisations, cultural institutions, and commercial entities have invested substantial funds creating digital assets to enable new research in the arts and

Light, Liturgy, and Art at the Monastery of Saint John in Müstair, Switzerland: A Software Demonstration

Ataoguz, Kirsten

kirstenataoguz@gmail.com

Indiana University-Purdue University Fort Wayne

In the early Middle Ages, solar observance shaped the art and architecture of Christian churches in various ways.

Medieval writers from across the Mediterranean related dramatic lighting effects to alignment with the rising sun on astronomically and liturgically significant days. Medieval

humanities. These assets have grown to include millions of items and petabytes of material covering all forms of content – manuscripts, monographs, maps, images, sound, and more. Unfortunately, scholars have been unable to fully exploit these digital assets. The supporting infrastructures are restrictive. The assets are distributed unevenly across organisations and systems. Access restrictions unpredictably limit where, how and who can use items.

This poster will outline a pathway to remedy this unacceptable state of affairs. It will explore the need for a simple-to-use infrastructure for digital scholarship. Built primarily using off-the-shelf technologies and services, we argue that such an interoperable infrastructure should, as far as possible, work like something the user already knows: it should allow the researcher to bring their own content, tools and creativity to a familiar environment. Where we envisage it differing from a local PC setup is by hosting otherwise difficult to obtain and too big to download digital content, offering the computational capacity required to quickly analyse big data using automated processes, and providing network services capable of robustly supporting digitally-driven research.

A key context of this proposed poster is research infrastructure developments around cloud, virtual and remote workflows. Notable among these are ongoing cyber-infrastructure work at the HathiTrust Research Centre¹ and the deployed cloud research infrastructure used by the European Bioinformatics Institute.² Whilst these observations and experiences point to a potentially crucial role for infrastructure in humanities research, we remain mindful of the robust critiques of recent digital humanities infrastructure projects. Quinn³ These critiques have highlighted how infrastructure development should not make strong assumptions about how researchers work, what tools they need, the sorts of problems that they will strive to solve, or even the specialised standards that they will employ. Our proposed pathway avoids these known problems by suggesting that researchers must be enabled to bring their own tools, work in whatever way they want, use any workflow, and address any sort of problem. We envisage this being achieved by infrastructure development that works with many digital content providers, supports a wide range of content types, and is embedded within arts and humanities research that uses a variety of data-driven methodologies. It would support growth in big data research in the arts and humanities using researcher appropriate standards and guidelines.

The informal, conversational setting of a poster session will prove a valuable opportunity to visit the key questions and problems around digital research infrastructure. These include:

- What are the benefits of scholars being able to use off-the-shelf technologies to work with big data across major content holders?
- How can these infrastructures enable transformative research?
- Do hybrid cloud infrastructures provide a sustainable approach to service provision?

Such infrastructure could establish the foundation for scholarly work with large scale content collections for years to come, enabling in turn transformative research that uncovers the value hidden in these digital assets and society to benefit from its investment. Such transformation requires leading-edge researchers, and eventually the majority of researchers, to adopt, learn and use new methods and techniques; to not just answer old questions in new ways but to arrive at new answers and to start asking entirely new questions as a consequence. This proposed infrastructure pathway aims to explore the next steps towards making this transformation a reality.

This poster builds on experience providing researchers with digital content. Scholars increasingly demand scalable access to large quantities of digital content – big data – that they can analyse using their own software and tools. Early on, the amounts of digital data were small; it was possible to provide copies or enable network downloads. With the growing volumes of big data, this is no longer plausible. Instead of moving hundreds of terabytes of data to researchers, we must allow researchers to bring their tools to the data. This is consistent

with changes in the broader IT landscape. We have established five principles to guide our pathway:

1. Keep it simple. Any new infrastructure should be simple to use and understand.
2. Lower the bar. Any new infrastructure should not expose or require users to understand new or complex technologies or processes. It should, as much as possible, work like something they already do
3. Bring your own tools. Users should be able to employ the tools that they already understand and work with. For example, if a researcher uses Mathematica for image analysis in her office, she should be able to use it on large collections of digital assets distributed across multiple content organisations.
4. Be creative. Users should be able to use data in creative, novel, unexpected ways. Many systems and infrastructures limit what users can do.
5. Start small and grow big. Users should be able to try things out; explore, experiment and debug; and then deploy on large content sets.

References

1. Beth Plale, Opportunities and Challenges of Text Mining HathiTrust Digital Library, Koninklijke Bibliotheek, 15 November 2013 www.hathitrust.org/documents/kb-plalehtrc-nov2013.pdf
2. Creating a Global Alliance to Enable Responsible Sharing of Genomic and Clinical Data, 3 June 2013 www.ebi.ac.uk/sites/ebi.ac.uk/files/shared/images/News/Global_Alliance_White_Paper_3_June_2013.pdf
3. Quinn Dombrowski, *What ever happened to Project Bamboo?*, DH2013

Neue Möglichkeiten der Arbeit mit strukturierten Sprachressourcen in den Digital Humanities mithilfe von Data-Mining

Bartz, Thomas

thomas.bartz@tu-dortmund.de

Technische Universität Dortmund, Institut für deutsche Sprache und Literatur, Germany

Beißenberger, Michael

Technische Universität Dortmund, Institut für deutsche Sprache und Literatur, Germany

Pöltz, Christian

Technische Universität Dortmund, Fakultät Informatik, Germany

Radtke, Nadja

Technische Universität Dortmund, Institut für deutsche Sprache und Literatur, Germany

Storrer, Angelika

Technische Universität Dortmund, Institut für deutsche Sprache und Literatur, Germany

1. Projekthintergrund: Ziele, Methoden, Ressourcen

Strukturierte Sprachressourcen (annotierte Textkorpora, Baumbanken, Wortnetze) bieten neuartige und attraktive Möglichkeiten, linguistische Fragestellungen an authentischen Sprachverwendungsdaten zu untersuchen und quantitativ auszuwerten (vgl. z.B. McEnergy et al. 2006, Lüdeling & Kytö 2008/2009). Infrastrukturprojekte wie CLARIN bieten flexible Werkzeuge an, um aus diesen Ressourcen Daten zu gewinnen und auszuwerten. Für sehr viele linguistische Forschungsfragen müssen die automatisch gewonnenen Ergebnisse allerdings noch weiter bearbeitet werden – gerade

wenn die Anwender nicht selbst Softwarelösungen für die Datenauswertung entwickeln können, sehen sie sich mit zeitaufwändigen, manuellen Routinearbeiten konfrontiert. Im Verbundprojekt *Korpus-basierte linguistische Recherche und Analyse mit Hilfe von Data-Mining (KobRA)* arbeiten Partner aus Informatik, Linguistik und Sprachtechnologie gemeinsam daran, die quantitative Auswertung strukturierter Sprachdaten zu verbessern und zu beschleunigen. Dazu werden im Rahmen korpusbasierter linguistischer Studien, die mit konkreten Forschungsaktivitäten der Projektbeteiligten in Verbindung stehen, Data-Mining-Verfahren (insbesondere Lernverfahren) im Zusammenspiel mit vorhandenen Sprachressourcen erprobt und angepasst. Die Verfahren operieren auf den Suchtrefferlisten bzw. auf großen Korpora und gehen über die reine Suche hinaus, indem sie die Suchergebnisse filtern, sortieren oder strukturieren sowie ggf. die weitere Aufbereitung der Daten für eine konkrete Fragestellung erleichtern. In unserem Vortrag stellen wir den Ansatz des Projekts vor (Abschnitt 2) und berichten über erste Ergebnisse (Abschnitte 3 und 4).

2. Projektarchitektur

Die Data-Mining-Verfahren des Projekts setzen auf der Infrastruktur der Sprachtechnologie-Partner auf. Es gibt einerseits eine Schnittstelle zu den linguistischen Anwendern und andererseits eine interne Schnittstelle zwischen der Data-Mining-Komponente und der Infrastruktur. Das Schaubild in Abbildung 1 verdeutlicht diese Verzahnung.

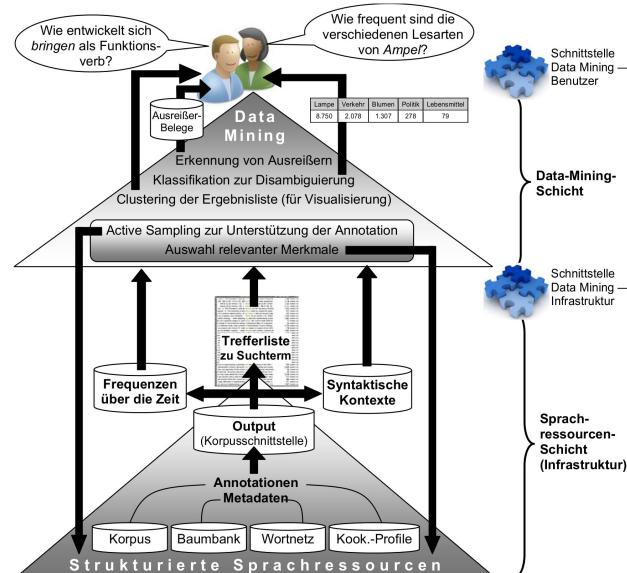


Fig. 1: Verzahnung und Schnittstellen zwischen den Projektkomponenten

Einige der im Projekt zu entwickelnden Lernverfahren werden direkt auf den Ergebnislisten (inkl. Annotationen und Metadaten) der von der Berlin-Brandenburgischen Akademie der Wissenschaften, dem Institut für deutsche Sprache (Mannheim) und dem Seminar für Sprachwissenschaft der Universität Tübingen bereitgestellten Sprachressourcen ausgeführt. Andere Verfahren operieren an der Schnittstelle zwischen der Data-Mining-Komponente und der Korpusinfrastruktur.

Zum Einsatz kommen bislang Verfahren der Klassifikation (z.B. Stützvektormethode (SVM), vgl. Joachims 2002) und des Clusterings (v.a. Topic Models, z.B. Latent Dirichlet Allocation (LDA), vgl. Blei et al. 2003; Tomanek & Morik 2011), die die automatische Bereinigung und Disambiguierung bzw. Klassifikation von Treffern (ggf. auf Basis einer möglichst geringen Menge intellektuell analysierter Treffer, z.B. mithilfe von Active Learning, vgl. Tomanek 2010, Tomanek & Morik 2011) ermöglichen. Um die Nutzer bei der Exploration verschiedener strukturierter Datenbestände zu unterstützen, werden auch innovative Formen der Visualisierung für typische sprachbezogene Forschungsfragen erprobt.

3. Fallstudien zu den Bereichen Lexikographie und Diachronische Sprachforschung

Erste Ergebnisse zum Nutzen von Data-Mining-Verfahren für konkrete korpusbasierte Forschungsvorhaben liegen bereits vor. Im Einzelnen wurden Verfahren für folgende Vorhaben angepasst und evaluiert:

- Studien zu deutschen Stützverbgefügen.
- Studien zur korpusgestützten lexikographischen Beschreibung von Wörtern mit mehreren Lesarten.

Ad a) Stützverbgefüge sind Konstruktionen aus einem prädiktiven Nomen und einem semantisch blassen Stützverb wie z.B. *Anwendung finden* oder *zur Anwendung kommen*. Im Rahmen eines Forschungsprojekts der Dortmunder Projektleiterin zur diachronen Entwicklung und Textsortenspezifität von Stützverbgefügen wurden erstmals große Korpusbestände aus unterschiedlichen Textsortenbereichen untersucht (vgl. Storrer 2013a). Weil die formbasierte Suche in den Korpora bislang keine Möglichkeit bietet, automatisch zwischen Vollverb- (*etw. finden*) oder Stützverbverwendungen (*Anwendung/Beachtung finden*) zu unterscheiden, mussten die ermittelten Suchtreffer manuell-intellektuell analysiert werden. Die dabei entstandenen annotierten Daten wurden im KobRA-Projekt genutzt, um ein automatisches Klassifikationsverfahren für Stützverben zu lernen. Bisher wurden SVM-basierte Klassifikationsverfahren (Stützvektormethode, vgl. Joachims 2002; als Merkmale wurden Kontextwörter und syntaktische Strukturen berücksichtigt) evaluiert, die aktuell auf Trefferlisten aus dem DWDS-Kernkorpus des 20. Jh. abhängig von Verb und Textsortenbereich eine Genauigkeit (Precision) von zwischen 70 und 87% sowie eine Ausbeute (Recall) von zwischen 36 und 80% erreichen. An einer Verbesserung der Ausbeute wird derzeit noch gearbeitet. Die Klassifikationsverfahren werden für den korpusgestützten Aufbau eines Wikis zu deutschen Stützverbgefügen genutzt.

Ad b) Als Ausgangspunkt für die Studien zur korpusgestützten lexikographischen Beschreibung von Wörtern mit mehreren Lesarten ist das Problem, dass strukturierte Sprachressourcen momentan noch nicht in semantisch disambiguierter Form vorliegen. Automatische Frequenzerhebungen beziehen sich deshalb immer nur auf Formeinheiten; für die lexikographische Arbeit ist man aber gerade auch an den Frequenzen zu einzelnen Lesarten homographer bzw. polysemer Wörter interessiert (z.B. für das Wort *Leiter*: *Sprossenstiege*, *Tonfolge*, *Verantwortlicher/Vorsteher*, *Energie übertragender Stoff*). Um Wörter wie Leiter adäquat beschreiben zu können, müssen korpusbasiert arbeitende Lexikographen bislang sämtliche Treffer zu einem Suchwort sichten (für *Leiter*: 6895 Treffer im DWDS-Kernkorpus des 20. Jh.); Werkzeuge zur automatischen Disambiguierung wären deshalb sehr hilfreich. Sie könnten auch statistische Analyse- und Visualisierungswerkzeuge verbessern (z.B. Kookurrenzanalysen, Wortverlaufsdiagramme), die bislang ebenfalls nicht zwischen Lesarten differenzieren. Aus diesem Grund werden im KobRA-Projekt Clusteringverfahren zur Partitionierung von Suchtrellerlisten nach Lesarten eines gesuchten Wortes evaluiert und angepasst. Beim Clustering von Trefferlisten aus dem DWDS-Kernkorpus des 20. Jh. zu den Wörtern *Leiter* und *zeitnah* (*zeitgenössisch*, *zeitkritisch* vs. *unverzüglich*) mithilfe von LDA-Topic-Models (vgl. Blei et al. 2003) konnten bislang F1-Werte (gleich gewichtetes Mittel zwischen Genauigkeit (Precision) und Ausbeute (Recall)) zwischen 74 und 78% erreicht werden. Dabei wurde die Partitionierung zunächst lediglich auf Basis der Kontextwörter (Bags-of-Words) vorgenommen. Aktuell wird auch der Nutzen weiterer Merkmale (Wortarten, Syntax, Textsorte, Erscheinungsdatum) erprobt.

4. Fallstudien zum Bereich Varietätenlinguistik / Internetbasierte Kommunikation

Die Kommunikation auf der Grundlage internetbasierter Kommunikationstechnologien und sozialer Medien stellt ein wichtiges neues Teilgebiet der Digital Humanities dar. Bei der

interpersonalen Kommunikation in Genres wie Online-Foren, Weblogs, Chats, Twitter oder sozialen Netzwerken finden sich Produkte schriftlicher Sprachverwendung, deren sprachliche Gestaltung an den Bedingungen dialogischer Kommunikation im sozialen Nähebereich orientiert ist. Typische Merkmale der *interaktionsorientierten Schreibhaltung* (Storrer 2013b), auf die die Orientierung an der Mündlichkeit in der schriftlichen internetbasierten Kommunikation (IBK) zurückgeführt werden kann, sind u.a. Phänomene geschriebener Umgangssprache wie etwa Verschmelzungs-/Allegroformen (*haste, biste, willste, machstes, isses, aufm, aufn*), Schwa-Elisionen (*ich schreib, ich mach, ich sag*), die Verwendung umgangssprachlicher Lexik (*moin, Maloche*) oder dialektal/regional gebundener Aussprachevarianten (*Oida wos wüst < Alter, was willst (du); wech < weg*) sowie die häufige Verwendung von Einheiten wie Interjektionen und Abtönungspartikeln. Darüber hinaus bilden sich in der schriftlichen internetbasierten Kommunikation sprachliche Mittel aus, die auf die Unterstützung der interaktiven schriftlichen Kommunikation am Nähepol optimiert sind. Typische Beispiele dafür sind Inflektive (*freu, lach, grübel, wirk, seufz*) und Inflektivkonstruktionen (*wildsei, malanmerk, bedenkenhab*), Emoticons sowie die Nutzung von Verfahren der Graphemiteration (*gaaaaaaaanz schlecht*) und der Großschreibung (*mathe mündlich? BRUTAL!!!*) für die graphische Nachbildung stimmlicher Kommunikationssignale.

Um die Besonderheiten der Schreibformen und sprachlichen Besonderheiten in der internetbasierten Kommunikation empirisch begründbar in einen sprach- und varietätengeschichtlichen Rahmen einordnen zu können, müssen Ausgangsbedingungen geschaffen werden, die einen Vergleich von Phänomenen konzeptioneller Mündlichkeit in internetbasierter Schriftlichkeit und dem Schreibgebrauch in historischen Korpora ermöglichen. Beim Aufbau von IBK-Korpora stellen sich derzeit noch viele Herausforderungen (vgl. z.B. Beißwenger & Storrer 2008, Storrer 2013b: Abschnitt 4), weil Verfahren und Standards, die sich für die Annotation von Textkorpora bewährt haben (Annotationsstandards, Metadatenschemata, Werkzeuge und Tagsets für die linguistische Analyse), nicht ohne Anpassungen für IBK-Korpora übernommen werden können.

Im KobRA-Projekt werden auf der Grundlage manuell annotierter Trainingsdaten Verfahren zur Klassifizierung und Disambiguierung auf die Behandlung von Phänomentypen (Verschmelzungen, Inflektive, Emoticons) trainiert, die in der Domäne typischerweise auftreten und die von Verarbeitungswerkzeugen, die auf den Umgang mit redigierten Texten trainiert sind, nicht angemessen behandelt werden können. Als Testbett für diese Verfahren dienen Daten aus verschiedenen im Aufbau befindlichen IBK-Korpora, die im Projekt zur Verfügung stehen – u.a. aus dem Wikipedia-Korpus am Institut für deutsche Sprache (Mannheim), dem Projekt „Deutsches Referenzkorpus zur internetbasierten Kommunikation“ (DeRiK, Beißwenger et al. 2013) sowie dem Dortmunder Chat-Korpus (Beißwenger 2013). Die Verfahren sollen in Arbeiten zur Anpassung von Werkzeugen für die automatische Wortartenannotation auf die Verarbeitung von IBK-Daten einfließen. Die Annotation erfolgt auf der Grundlage einer erweiterten Version des STTS-Standards für das POS-Tagging deutscher Sprachdaten, in dessen Erarbeitung die Projektbeteiligten involviert sind (Bartz et al. 2013). Sie ist abgestimmt auf Aktivitäten zur Erarbeitung eines Standards für die Strukturannotation von IBK-Korpora im Rahmen der *Text Encoding Initiative* (TEI).

References

- Bartz, T., Beißwenger, M., Storrer A.** (2013): *Optimierung des Stuttgart-Tübingen-Tagset für die linguistische Annotation von Korpora zur internetbasierten Kommunikation: Phänomene, Herausforderungen, Erweiterungsvorschläge*. Journal for Language Technology and Computational Linguistics (Themenheft „Das STTS-Tagset für Wortartentagging – Stand und Perspektiven“).
- Beißwenger, M.** (2013): *Das Dortmunder Chat-Korpus*. Zeitschrift für germanistische Linguistik 41/1, 161–164. (Erweiterte Fassung online: <http://tinyurl.com/chatkorpus>).
- Beißwenger, M., Ermakova, M., Geyken, A., Lemnitzer, L., Storrer, A.** (2012): *A TEI Schema for the Representation of Computer-mediated Communication*. Journal of the Text Encoding Initiative (jTEI), Issue 3, jtei.revues.org/476 (DOI: 10.4000/jtei.476).
- Beißwenger, M., Ermakova, M., Geyken, A., Lemnitzer, L., Storrer, A.** (2013): *DeRiK: A German Reference Corpus of Computer-Mediated Communication*. In: Literary and Linguistic Computing. tinyurl.com/derik-llc (DOI: 10.1093/llc/fqt038).
- Beißwenger, M. and Storrer, A.** (2008): *Corpora of Computer-Mediated Communication*. In Lüdeling, A. und Kytö, M. (eds), *Corpus Linguistics. An International Handbook*. Volume 1. Berlin, New York: de Gruyter (Handbücher zur Sprache und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science 29.1), pp. 292–308.
- Blei, D. M., Ng, A. Y., Jordan, M. I.** (2003): *Latent dirichlet allocation*. Journal of Machine Learning Research 3, pp. 993–1022.
- Geyken, A.** (2007): *The DWDS corpus: a reference corpus for the German language of the 20th century*. In Fellbaum, C. (ed), Idioms and collocations. Corpus-based linguistic and lexicographic studies. London: Continuum, pp. 23–40.
- Joachims, T.** (2002): *Learning to Classify Text Using Support Vector Machines*. Dissertation. Dordrecht: Kluwer.
- Krenn, B., Erbach, G.** (1994): *Idioms and support verb constructions*. In Nerbonne, J., Netter, K., Pollard, C. (eds), *German in Head-Driven Phrase Structure Grammar*. Stanford: CSLI publications, pp. 365–395.
- Langer, S.** (2004): *A linguistic test battery for support verb constructions*. Linguisticae Investigationes 27 (2), pp. 171–184.
- Lüdeling, A. and Kytö, M. (eds)** (2008/9): *Corpus Linguistics. An International Handbook*. 2 Bände. Berlin, New York: de Gruyter.
- Morik, K., Kaspari, A., Wurst, M., Skrzynski, M.** (2012): *Multi-objective frequent termset clustering*. Knowledge and Information Systems 30 (3) (DOI:10.1007/s10115-011-0431-3), pp. 715–738.
- McEnergy, T., Xiao, R., Tono, Y.** (2006): *Corpus-Based Language Studies*. An Advanced Resource Book (Routledge Applied Linguistics). London, New York: Routledge.
- Storrer, A.** (2007): *Corpus-based investigations on German support verb constructions*. In Fellbaum, C. (ed), Idioms and collocations. Corpus-based linguistic and lexicographic studies. London: Continuum Press, pp. 164–188.
- Storrer, A.** (2013a): *Variation im deutschen Wortschatz am Beispiel der Streckverbgefüge*. In Deutsche Akademie für Sprache und Dichtung; Union der deutschen Akademien der Wissenschaften (eds), Reichtum und Armut der deutschen Sprache. Erster Bericht zur Lage der deutschen Sprache. Berlin/New York: de Gruyter, pp. 171–209.
- Storrer, A.** (2013b): *Sprachverfall durch internetbasierte Kommunikation? Linguistische Erklärungsansätze – empirische Befunde*. Sprachverfall? Dynamik – Wandel – Variation. Jahrbuch des Instituts für Deutsche Sprache 2013.
- Tomanek, K.** (2010): *Resource-aware annotation through active learning*. Dissertation, TU Dortmund.
- Tomanek, K., Morik, K.** (2011): *Inspecting Sample Reusability for Activ Learning*. JMLR Workshop and Conference Proceedings 16, pp. 169–181.
- Das Verbundprojekt wird vom Bundesministerium für Bildung und Forschung (BMBF) seit Herbst 2012 im Rahmen des Programms „eHumanities“ gefördert. Informationen zu den Projektbeteiligten und den Ergebnissen unter: www.kobra.tu-dortmund.de.
- Engl. „support verb constructions“ (vgl. u.a. Langer 2004, Krenn & Erbach 1994, Storrer 2007/2013).
- Zu den Korpora im Projekt „Digitales Wörterbuch der deutschen Sprache“ (DWDS) vgl. Geyken (2007).
- Relevant sind hier v.a. die Aktivitäten der Special Interest Group „Computer-Mediated Communication“ (<http://www.tei-c.org/Activities/SIG/CMC/>); ein erster Entwurf für ein TEI-Schema für die Annotation von Genres internetbasierter Kommunikation ist in Beißwenger et al. (2012) beschrieben.

Mapping Colonial Americas Publishing Project

Bauer, Jean

jean_bauer@brown.edu
Brown University

Egan, James

james_egan@brown.edu
Brown University

In 1524, less than thirty years after arriving on American soil, Europeans opened the New World's first print shop. By the time creole nationalist movements up and down the two continents were in full swing three hundred years later, the Americas had a thriving print trade with presses spread across two continents.¹ *The Mapping Colonial Americas Publishing Project* (cds.library.brown.edu/projects/mapping-genres) aims to visualize New World printing over geographic space and across literary genres from European contact to 1800. Our poster will illustrate the progress we have made thus far in our efforts to visualize what kinds of works were published where in the Americas before 1800 and how these printing patterns changed over time.

Scholars who study the print trade in the Americas before 1800 have long known that the types of material printed in Europe's American colonies varied according to region and historical period.² Scholars have yet to visualize data culled from library catalogs to illustrate the dramatic differences in the kinds of genres published in different parts of the Americas and the way those generic patterns changed over time. Mapping Colonial Americas Publishing aims to put these print histories on display.

By representing the generic differences in the history of print in both North and South America, Mapping Colonial Americas Publishing allows scholars, students, and others to compare generic patterns across languages, cultures, and nations.

Scholarship on print in pre-1800 North and South America has, until recently, focused either on English or Spanish materials. Our project fits with recent trends toward Hemispheric study of the colonial Americas, and our poster will highlight this hemispheric approach to colonial print histories.³

Early visualizations based on data from the Brown University Library, especially the Universities rare book libraries, the John Carter Brown and the John Hay, as well as the American Antiquarian Society have revealed the complexity of library catalogs as a digital-print hybrid genre. Literary scholars constantly use library catalogs to locate materials, but rarely think about the conceptual structures or technical possibilities (and limitations) inherent in the data collected by librarians and archivists over the centuries. Library cataloging systems were designed to group, retrieve, and display records, and provide powerful tools for name authority, but individual records are created by hand and many crucial fields for book history (title, imprint, place of publication) contain long sections of multilingual, free form text, which resist systematic parsing to extract spatial or temporal information.⁴

Rare book catalogers esteem fidelity to the original title page, including spelling variants and latinized proper nouns (ex. Yale University Press giving its place of publication as "Novo Porto" instead of New Haven). This makes an individual catalog entry a rich source for work on the cultural history of reading and the materiality of the book trade, but presents real challenges for collating these entries to perform geospatial analysis or genre comparison over one hundred years of changing best practices.

Library catalog data has become a topic of renewed interest in Digital Humanities with the rise of large scale data repositories such as the HathiTrust and the Digital Public Library of America, but these projects have focused their efforts on combining millions of records into interoperable and searchable data structures.⁵ *The Mapping Colonial America's Publishing Project* is dealing with a much smaller dataset (~140,000 catalog records), produced at Brown University, which allows us to delve into the intricacies of the cataloging

and to remodel the data to suit our own research interests and the needs of the Brown University community.

The project will have two major end products: a gazetteer of normalized and geolocated places of publication in the Americas before 1800 and a series of data visualizations designed to help scholars and students explore printed material on the axes of genre, place of publication, date of publication, and format. Following on our initial dives we are refocusing our efforts to build a visual portal into the printed materials from the colonial americas physically held by Brown University. A smaller data set will allow us to do the more intensive data cleaning required for Thiessen Polygon analysis (locations) or data mining (titles, subject headings). Using local data will also allow us to work with our catalogers and verify our analysis against the physical objects as necessary. This project is using open source tools, including OpenRefine (<http://openrefine.org>) for data cleaning and the D3.js data visualization library (<http://d3js.org>). All the code and cleaned library data produced by this project is available on github under the MIT license for anyone to download and explore.

References

1. For a history of the book trade in colonial British America, see Hugh Armory and David D. Hall, eds., *A History of the Book in America. Volume One. The Colonial Book in the Atlantic World* (Cambridge: Cambridge UP, 2000); Hellmut Lehman-Haupt, *A History of the Selling and Making of Books in the United States* (New York : R.R. Bowker Co., 1951), 2nd. ed., rev. and enl. ed. and Lawrence C. Wroth, *The Colonial Printer* (NY: Dover, 1965). For a history of the book trade in colonial Spanish America, see Julie Greer Johnson, *The Book in the Americas: The Role of Books and Printing in the Development of Culture and Society in Colonial Latin America* (Providence, RI: John Carter Brown Library, 1988) and Hensely C. Woodbridge and Lawrence S. Thompson, *Printing in Colonial Spanish America* (Troy, NY: Whitson, 1976).
2. Scholarship that focuses on the regional dimension of colonial print cultures include Armory and Hall, Trish Loughran, *The Republic in Print: Print Culture in the Age of U.S. Nation Building, 1770-1870* (NY: Columbia, 2007). Michael Warner, *The Letters of the Republic: Publication and the Public Sphere in Eighteenth-Century America* (Cambridge, MA: Harvard UP, 1990), and Wroth.
3. For examples of recent studies that argue for a hemispheric approach to literary and historical approaches to the study of colonial European cultures in the New World, see Ralph Bauer, *The Cultural Geography of Colonial American Literatures: Empire, Travel, Modernity* (Cambridge: Cambridge UP, 2003) and Anna Brickhouse, *Transamerican Literary Relations and the Nineteenth-Century Public Sphere* (Cambridge: Cambridge UP, 2004).
4. Two excellent recent explorations of cleaning catalog data are Mia Ridge, "Mia Ridge explores the shape of Cooper-Hewitt collections," Cooper-Hewitt Labs, June 19, 2012. labs.cooperhewitt.org/2012/exploring-shape-collections-draft/ and Lincoln Mullen, "Quantifying the American Tract Society: Using Library Catalog Data for Historical Research," Religion in American History, August 1, 2013. usreligion.blogspot.com/2013/08/quantifying-american-tract-society.html
5. For example, A Preservation Infrastructure Built to Last: Preservation, Community, and HathiTrust. UNESCO The Memory of the World in the Digital Age: Digitization and Preservation, Vancouver, British Columbia, September 26-28, 2012. (September 2012) - Jeremy York. On the visualization side, the DPLA is experimenting with a number of large scale data visualizations, of which an ever growing list can be found at dp.la/apps.

Data Curation Nightmare: Migrating VM/CMS to GNU/Linux in 2 weeks

Bauman, Syd

s.bauman@neu.edu

Northeastern University

DiCamillo, Peter

Peter_DiCamillo@Brown.edu

Brown University

Data Curation Nightmare: Migrating VM/CMS to GNU/Linux in 2 weeks

Data curation has been described as "activities [that] enable data discovery and retrieval, maintain quality, add value, and provide for re-use [of digital assets] over time"¹. Data migration is a major and necessary (but not sufficient) part of long-term data curation.

At the time of the project described, the authors had worked at Brown University in the computing arena since 1985 and 1979 respectively. Up until the mid-1990s this work was almost exclusively on or related to the IBM VM mainframe system, but from the mid-1990s onward work and computer usage gradually migrated off the mainframe to Unix systems, GNU/Linux systems, and desktop Macintosh computers. By early 2009 almost no one at Brown was using the mainframe.

However, the authors were each responsible for a significant quantity of data and programs on the mainframe, including various system utilities, application programs each had written, and almost all of the historical digital assets of the Women Writers Project. In 2009, Brown decided to retire the mainframe.

Thus the authors migrated a large quantity of digital data, comprising widely ranging kinds of files, from a legacy IBM VM system to a modern GNU/Linux system. This was a significant undertaking, not just because of the quantity of data (~ 3 GiB), but because a VM system is *very* different than a GNU/Linux system. Differences include:

- files are named with a different naming convention
- files are stored in one of several internal formats, none of which is like that on a GNU/Linux system
- files are grouped in a flat, rather than hierarchical, system
- the underlying character encoding is not only not Unicode, it's not ASCII -- it is EBCDIC, which is quite different
- some files have been automatically "packed" or compressed by the system
- many files had been put into compressed archives (not unlike ZIP files) that could not be read on a GNU/Linux system
- many files were stored not on the mainframe itself, but on tapes that we could not read without the mainframe
- some bits of metadata on the mainframe have no counterpart in GNU/Linux

Moreover, this project had to be executed under very tight time constraints without administrative support.

This particular project was further complicated because it was envisioned not just as a curation of textual data (e.g., program source code, text formatting files, or SGML files), but as preservation of executable programs as well. One goal was to be able to move this data to a different IBM VM system without loss of any crucial information, such that the programs could still be run. It is worth noting that we were not archiving these materials in the analog archivist's sense -- i.e., we didn't decide what material to keep and what to discard, we basically kept it all.

While it must be the case that others have migrated data off of IBM VMsystems, the authors are not aware of any similarly ambitious projects. Having no blueprint to follow, the authors had to invent a method for transferring data from VM to GNU/Linux in a manner that would both keep all of its original properties intact (so that it could subsequently be moved to another VM system), and simultaneously permit direct use of cross-platform data (e.g., source code, text formatting documents, JPEGs, etc.). This method involved writing at least 5 separate programs, including a 1751 line-long C program that can be used to extract files from VM "DISK DUMP" format files(essentially disk images) that have been transferred to a GNU/Linux system. (In theory this would work just as well on a Mac OS X system, and perhaps even on a Windows system.)

This paper treats this project as a case study, interesting both because it describes what many younger DHers would think of as a foreign or archaic system (or both) that was in heavy use at Brown a mere 20 years ago, and is still in use today (as of 2013-11-01, the latest version was released 2013-07-23), and because it gives evidence to how significant a problem migration can present.

Notes

Moreover, this project had to be executed under very tight time constraints without administrative support.

This particular project was further complicated because it was envisioned not just as a curation of textual data (e.g., program source code, text formatting files, or SGML files), but as preservation of executable programs as well. One goal was to be able to move this data to a different IBM VM system without loss of any crucial information, such that the programs could still be run. It is worth noting that we were not archiving these materials in the analog archivist's sense -- i.e., we didn't decide what material to keep and what to discard, we basically kept it all.

While it must be the case that others have migrated data off of IBM VMsystems, the authors are not aware of any similarly ambitious projects. Having no blueprint to follow, the authors had to invent a method for transferring data from VM to GNU/Linux in a manner that would both keep all of its original properties intact (so that it could subsequently be moved to another VM system), and simultaneously permit direct use of cross-platform data (e.g., source code, text formatting documents, JPEGs, etc.). This method involved writing at least 5 separate programs, including a 1751 line-long C program that can be used to extract files from VM "DISK DUMP" format files(essentially disk images) that have been transferred to a GNU/Linux system. (In theory this would work just as well on a Mac OS X system, and perhaps even on a Windows system.)

This paper treats this project as a case study, interesting both because it describes what many younger DHers would think of as a foreign or archaic system (or both) that was in heavy use at Brown a mere 20 years ago, and is still in use today (as of 2013-11-01, the latest version was released 2013-07-23), and because it gives evidence to how significant a problem migration can present.

References

1. Cragin, Melissa H.; Heidorn, P. Bryan; Palmer, Carole L.; Smith, Linda C. (2007), *An Educational Program on Data Curation*; poster session presented at ACRL STS 2007.

www.ala.org/ala/mgrps/divs/acrl/about/sections/sts/conferences/posters07.cfm
hdl.handle.net/2142/3493

The Open Philology Project at the University of Leipzig**Baumgardt, Frederik**baumgardt@informatik.uni-leipzig.de
University of Leipzig, Germany**Berti, Monica**monica.berti@uni-leipzig.de
University of Leipzig, Germany**Celano, Giuseppe**celano@informatik.uni-leipzig.de
University of Leipzig, Germany**Crane, Gregory R.**crane@informatik.uni-leipzig.de
University of Leipzig, Germany**Dee, Stella**

dee@informatik.uni-leipzig.de
University of Leipzig, Germany

Foradi, Maryam
maryam.foradi@uni-leipzig.de
University of Leipzig, Germany

Franzini, Emily
efranzini@informatik.uni-leipzig.de
University of Leipzig, Germany

Franzini, Greta
franzini@informatik.uni-leipzig.de
University of Leipzig, Germany

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

The Open Philology Project (OPP) at the University of Leipzig aspires to re-assert the value of philology in its broadest sense and has been designed with the hope that it can contribute to any historical language that survives within the human record. It includes three different yet interdependent tasks:

(1) **Open Greek and Latin Project (OGL)** : OGL is currently collecting and scanning editions of classical texts in an effort to build the largest and most comprehensive open-source library of classical philology to date, concurrently contributing to the expansion of Google Books. Where existing corpora of Greek and Latin have generally included one edition of a work, the OGL corpus is designed to manage multiple, copyright-free editions and translations.

The digitization workflow involves OCR, correction and encoding in EpiDoc-compliant XML. The large volume of data we aim to generate requires significant computational power and task management, thus entreating a partnership with two Data Entry companies who carry out each operation under the supervision of the Leipzig team. While performed by our contractors, OCR correction is facilitated and partly automated thanks to a proofreading tool jointly developed by Leipzig, Mount Allison University and the CNR (Bruce Robertson of Mount Allison University, Canada, and Federico Boschetti of the CNR, Italy). Works currently under conversion include, amongst others, the *Patrologia Latina*, the *Patrologia Graeca*, the *Commentaria in Aristotelem Graeca*.

Moreover, Leipzig has established international collaborations aiming at creating open-source, curated collections and electronic editions of Greek and Latin literature. Editorial projects include the *Digital Fragmenta Historicorum Graecorum* project, *Digital Athenaeus*, and *Bibliotheca Aeschylea*. Furthermore, collaborations with Croatia, Bulgaria and Georgia will yield machine-actionable versions of translations of classical literature in these languages, thus opening-up research into less-explored textual heritage.

(2) **Historical Languages e-Learning Project (eLP)** : the development of dynamic textbooks that use richly annotated corpora to teach the vocabulary and grammar of texts that learners have chosen to read, and at the same time engage users in collaboratively producing new annotated data. eLPs is developing computationally customized learning materials for historical languages, beginning with Ancient Greek. The text selected for the pilot is the *Pentecontaetia*, part of *Thucydides' History of the Peloponnesian War*. Users learn through active engagement with the text and through the contribution of their own annotations. Future work will extend the system to accommodate other corpora.

At the core of eLP lies increasing the accessibility and enjoyability of the morphosyntactic and semantic annotation of text (e.g. treebanking), including that deriving from the OGL corpus. The creation of such a richly annotated and searchable text repository will serve a variety of purposes, including research in philology, Natural Language Processing (NLP), historical linguistics, and second language acquisition (SLA).

The production of automated queries to support this dynamic, customized, and localized interface relies upon the backend storage of complex textual data. The chosen graph model meets the broad requirements of the e-Learning application while retaining features of the real world objects represented by the data. The absence of schemas within graph databases enables extensibility, while maintaining a stable experience for

users through the use of REST APIs.³ The web interface takes the data and adapts its presentation to individual needs and access devices. HTML5, CSS3, and responsive technologies provide an appropriate experience to users regardless of how they access the system, while templating systems allow for resources that are structurally accessible via any first language.

(3) **Open Publications and Data Revenue Models** : OPP is establishing a new model of scholarly publication in a born digital environment. Such a task is accomplished through *Perseids*, which is a collaborative platform for annotating TEI XML documents in Classics, including inscriptions and manuscripts. The main publication model within the OPP is the *Leipzig Open Fragmentary Texts Series*, whose goal is to establish open editions of ancient works that survive through quotations and re-uses in later texts. Such editions are fundamentally hypertexts and the effort is to produce a dynamic infrastructure for a full representation of relationships between sources, quotations, and annotations about them.

With open data meaning by definition free access for all users, the OPP team has already begun thinking of ways for it to be financially sustainable for years to come. The team intends to devise business models to sustain and maintain distributed open source learning and discourse. The core principle is to move away from charging for monopoly access to data, to charging instead for services that allow users to identify, analyze and then contribute to increasingly complex open data, with services for faculties, students and for the interested public set at recognized and affordable price points.

References

Bruce Robertson of Mount Allison University, Canada, and **Federico Boschetti** of the CNR, Italy.

Digital Fragmenta Historicorum Graecorum (DFHG): www.dh.uni-leipzig.de/wo/open-philology-project/the-leipzig-open-fragmentary-texts-series-lofts/digital-fragmenta-historicorum-graecorum-dfhg-project ; *Digital Athenaeus* (with the University of Nebraska and the Perseus Digital Library), *Bibliotheca Aeschylea* (with researchers based in Leipzig and in Italy).

For more information about REST APIs, see:
en.wikipedia.org/wiki/Representational_state_transfer
(Accessed: 4 March 2014).

What's in a Discipline? Research Practices, Use of Tools and Content in the Humanities and Social Sciences - The web-based questionnaires of EHRI and Europeana Cloud.

Benardou, Agiatis
Digital Curation Unit, ATHENA R.C., Greece

Papaki, Eliza
Digital Curation Unit, ATHENA R.C., Greece

Chatzidiakou, Nephelie
Digital Curation Unit, ATHENA R.C., Greece

This poster reports on work conducted during 2010-2012 in the context of *EHRI – the European Holocaust Research Infrastructure*, as well as work in progress in the context of *Europeana Cloud - Unlocking Europe's Research via the Cloud*. Its purpose is to investigate any differentiations between the research practices of humanists and social scientists as identified within the User Requirements work conducted in the context of those two EU Research Infrastructure Projects (Benardou et.al. 2013, Benardou et.al. 2010), by demonstrating the points of divergence and convergence of humanists and

social scientists with regard to their scholarly research activities in a concise and illustrated format.

In the context of *EHRI* the Digital Curation Unit, "ATHENA" R.C. (DCU) was responsible for the identification, modeling and formalization of the requirements of *EHRI* users – largely text-based humanists but also social and political scientists. To this end, DCU identified and analysed scholarly research practices and focused on the use of archival materials in the area of Holocaust Studies, as well as scholarly research practices in the digital domain and how these might support and enhance research in Holocaust Studies, in order to create a set of data and functional requirements based upon the analysis of scholarly research practices. The quantitative part of this research, complementary to a series of semi-structured interviews with Holocaust researchers, consisted of an online questionnaire survey which covered the relative use of different kinds of digital and analog resources, the perceived importance of specific information activities used by researchers (covering the span from information seeking to collaboration, including entry points), perceptions towards sharing and trustworthiness of resources, computer/device use and work location, and demographic/control variables such as country of residence, researcher status, expertise in archival research methods etc. 82,28% of the respondents of this survey were Humanists, predominantly historians, while the rest came from the Social Sciences.

Within *Europeana Cloud*, DCU is leading the Workpackage responsible for the improvement of the understanding of digital tools, research processes and content used in the Humanities and Social Sciences, thus informing the development of tools and content strategy in *Europeana Cloud*. To this end, amongst other user-centred approaches such as a series of Expert Fora, DCU designed a Research Communities Web Survey to analyze digital research practices, tools and content to gather evidence-based data from the Humanities (75,38%) and Social Sciences research community (24,62%), focusing in particular on the potential use of content from *Europeana* and *The European Library*.

Humanistic and Social Sciences have by and large been perceived as two associated fields for which often the study of scholarly work adopts related if not identical strategies. Significant examples include the work of Ellis' team, who identified six common processes across disciplines spearheaded by qualitative work, fuelled by grounded theory research on research communities across the Social Sciences and the Humanities. Moreover, in a more recent study of the University of Washington (2005) on the use of digital sources, researchers in these two disciplines were perceived as comparable with regards to their approaches and the methods they adopted. In this light, similarities rather than discrepancies were stressed. Such an approach is indeed largely reasoned, given the fact that the Humanities and Social Sciences broadly share methods and objects, and are obviously closer to each other when compared to, e.g., the Physical Sciences. However, many scientific papers on research information behaviour have differentiated the two fields. For instance, an interesting survey conducted by the British Academy on e-resources for research in the Humanities and Social Sciences (2005) stresses the different research approaches across the two fields. On discussing the questionnaire and the recorded answers, the authors distinguish the views expressed by humanists and social scientists, thus reaching conclusions on their distinct characteristics with regard to digital content.

It seems therefore interesting to investigate research activities in the Humanities and Social Sciences separately yet comparatively, in order to define whether there could be any reasonable ground for differentiation in the design and implementation of future infrastructures. To this end, we will be looking into issues addressed in both *EHRI* and *Europeana Cloud* online questionnaires, such as the use of specific tools and services, the research activities in which the users engage, the content as well as the properties of the resources favored by the users and the degree of agreement regarding specific statements concerning the overall research process, in an attempt to trace, identify and highlight similarities and differences on issues of workflow, concurrency, microactivities relating to information seeking behavior, aiming at mapping out

research practices at such granular level as to gather detailed information on current research practices among different user groups engaged in the Humanities and Social Sciences, which could act as a reference point and information resource for the formulation of data requirements and functional specifications for future infrastructures.

References

- Bass, A., Fairlee J., Fox K. & Sullivan J. (2005). *The Information Behavior of Scholars in the Humanities and Social sciences*. University of Washington.
- Benardou, A., Constantopoulos P. & Dallas C. (2013). *An Approach to Analysing Working Practices of Research Communities in the Humanities*. International Journal of History and Arts Computing, 7.1-2, 105-127.
- Benardou A., P. Constantopoulos, C. Dallas & D. Gavrilis (2010). *Understanding the information requirements of arts and humanities scholarship: Implications for digital curation*. International Journal of Digital Curation5, no. 1
- Ellis, D. (1993). *Modeling the information-seeking patterns of academic researchers: A grounded theory approach*. The Library Quarterly, 63(4), 469-486.
- Jones, K., and Bennett, R., "E-resources for research in the Humanities and Social sciences: A British Academic Policy Review", British Academy, April 2005, pp. 1-116.

The CENDARI Project: A user-centered 'enquiry environment' for modern and medieval historians

Benes, Jakub
j.benes@bham.ac.uk
University of Birmingham

O'Connor, Alex
Alex.OConnor@scss.tcd.ie
Trinity College Dublin

Dimara, Evangelia
evanthia.dimara@gmail.com
INRIA - Institut national de recherche en informatique et en automatique

1. Introduction

The Digital Humanities remains an exotic garden to many historians. While software developers have focused on sophisticated analytical tools that require large datasets and pointed research questions, historians often consider themselves unready to use such tools or regard them as superfluous once they have gathered and organized sufficient research data. To many, digitization projects seem too narrowly conceived to represent disciplinary breakthroughs, in part because they typically neglect archival sources. Moreover, immense national and institutional asymmetries exist in efforts to further digital history.

The CENDARI project overcomes some of these constraints. On the most basic level, it is integrating data and metadata from archives, libraries, and museums across Europe relevant to the project's two historical domain test cases: Medieval culture and World War I. In order to further transnational and comparative research, and to overcome entrenched historiographical and digital asymmetries, the project includes eastern and southern European repositories ('hidden archives' to many historians) along with the more visible western European institutions.¹

From a computer science perspective, the relevant data is dizzyingly heterogeneous in terms of languages, formats, level of granularity, completeness, encoding standards, annotation schemes, etc. Therefore CENDARI has implemented a capacious approach to data integration and curation based on the concepts of 'data space' and 'blackboard'. This will produce

a flexible and interactive digital ecosystem, underpinned by various ontologies, that enables collaborative research using a variety of digital tools. Cooperation with the European digital humanities infrastructure DARIAH will ensure the ecosystem's sustainability.

Historians will be able to access data by pursuing their own research projects through a dynamic user interface. While the enquiry environment is focused on the initial, exploratory phases of research, it will go beyond "search and retrieval." Historians will be able to analyze data with the help of sophisticated data mining and visualization tools; they will be able to upload their own research to a personal research space, and they will be able to curate and exchange data with other researchers through annotations, tags, semantic links, and other tools. Project partners have developed this enquiry environment based on interactive participatory design sessions, domain specific "use cases", and two domain-specific "prototype projects," all designed to integrate the user's perspective while the research infrastructure is built.

CENDARI incorporates archival data, and creates a research space where users can see projects through from finding and organizing sources to analyzing and sharing data with sophisticated tools. The project overcomes the national 'siloes' of digitization efforts and historical inquiry. Perhaps above all, it may help open digital history to the majority of professional historians, representing a major breakthrough in *digital cultural empowerment*.

1.2. Methodology

Approach to data

How can CENDARI help users answer questions they did not know they wanted to ask, and how can these users then be helped to record and share the process and results of those questions? The CENDARI project offers a unique opportunity to demonstrate "serendipity through heterogeneity". There is already an enormous number of web-based tools and projects which offer the web browser digital access to archives and collections. CENDARI will not attempt to become a "big data" repository for all of them. Instead, the project should recognize that the value of scholarship is in the interlinking of different concepts, objects, collections and content to highlight insights that are not otherwise obvious. This is among the primary goals of the project: to foster serendipity in research processes as well as to support auditable, traceable research trails.

CENDARI data are heterogeneous in the origin of their sources, formats, metadata profiles, type of content they hold, methods of acquisition or creation and distribution rights pertaining to them. In some cases, data will be stored within CENDARI, such as data produced within the context of the CENDARI Archival Directory (as metadata manually edited or coming from a particular repository with links to the original sources); in other cases these will have a more transient character, e.g. if based on a search results retrieved from external system.

A design goal of the CENDARI data infrastructure is to build an interoperable data platform, overcoming various data siloes and leveraging the potential of already existing platforms and their existing data services "below the level of work."² Additionally, CENDARI aims to reach a more detailed level of data granularity as the basis for real scholarly work and employ services that support knowledge discovery, organization and sharing.

We address the aspect of infrastructure development that embraces data diversity, i.e. the "data soup," and takes an incremental approach to the data integration, based on the concept of Dataspaces,³ SOA⁴ and an adapted "Blackboard" model approach⁵, while employing information extraction, NLP tools and statistical methods in order to build infrastructure components for historical research.

Approach to the Virtual Research Environment

Researcher involvement was seen as a key element in all aspects of the technical development. The partners in charge of defining the system architecture and designing the User Interface (UI) employed several methods, such as video brainstorming sessions for the creation of mockups, for

understanding the user requirements and methodological needs of the target users: World War I historians and medievalists. Project historians also analyzed their own research methods, and began communicating them to technical specialists, by creating a number of scenarios drawing on concrete research inquiries. The two most detailed of these were selected to serve as "prototype projects" that constituted both real research endeavors and a means of defining the technical functionalities of the enquiry environment.

The iterative design process revealed strong user interest in a VRE centered on an advanced note-taking environment with links to the CENDARI data space, continuously enriched by historians' notes. This result came from the conjunction of interesting findings: all the historians take notes, either on paper, in digital form, or both. From their notes, they try to resolve people (who is that person?), places, dates, artifacts, events, and organizations, among other entities. This resolution leads them to search for related entities (e.g. the family of that person, the archive holding information related to that event), until they reach a point where they have a clearer picture of a situation, or they give up for lack of information. Relating entities is a complex task not well supported by existing digital environments. Historians would like to search in their colleagues' notes for hints, but are opposed to sharing their own notes by fear of being "scooped". To avoid the problem, the VRE allows searching in entities contained in notes without disclosing the contents of the notes in their entirety. Brainstorming with historians revealed that they would accept sharing the entities only (with some control). Therefore, note-taking from multiple historians weaves a network of entities, creates a resource that facilitates connecting information, and allows asking appropriate colleague historians for help.

Our primary design goal is a technology that does not interrupt historians' workflow. We propose a smooth and on-demand integration of intelligent tools, like the entity recognizer, so the researcher has full control of his project.

In order to make the VRE easy to learn, our design mimics the traditional historian's physical workspace. Based on the participatory design insights, the VRE aims to interpret the affordances of the historian's personal library, note taking, entity highlighting, annotations or work organization to digital tools. The notion of "affordance" here implies that the appearance of the tool reveals a part of its functionality to the user. Once the researcher is able to accelerate his working rate in VRE, we enrich the workflow with individualized visualizations based on the user scenario's queries. Our design approach is based on the researcher's daily routine. We use an agile software development methodology to allow quick adaptation of the system to historians' needs.

In an era in which the digital can drive much scholarly innovation, this note-taking environment meets and serves the needs of historians, who generally keep a traditional research diary or notebook. At the same time, it seems to foster new research approaches and new attitudes towards the organization and use of archival sources. Seen from a user/researcher's perspective, the note-taking environment could therefore be an interesting platform for both organizing existing data and notes, and for envisaging new research directions. The concept of selective sharing represents a new opportunity for experiencing research work in a selected and collaborative environment which, when properly understood and used, might boost the potential of archival work accomplished across different countries.

References

1. In this, digitization initiatives have generally reflected supply and opportunity rather than demand. See Melissa Terras, "Digitization and digital resources in the humanities" in Claire Warwick, Melissa Terras, and Julianne Nyhan, eds., *Digital Humanities in Practice* (London, 2012).
2. P. Edwards, S. Jackson, G Bowker and C Knobel, *Understanding Infrastructure: Dynamics, Tensions and Design* hdl.handle.net/2027.42/49353 , last accessed 21 May 2013.
3. Norbert Antunes, Tatiana Malyuta, and Suzanne Yoakum Stover, *A Data Integration Framework with Full*

Spectrum Fusion Capabilities, Presented at the Sensor and Information Fusion Symposium, Las Vegas, NV, August 2009, 2-3.

4. Krafzig, Dirk; Karl Banke, Dirk Slama: Enterprise SOA: Service-Oriented Architecture Best Practices, (New Jersey, 2004), ISBN 978-0-13-146575-6.

5. Lee D. Erman, Frederick Hayes-Roth, Victor R. Lesser, and D. Raj Reddy. *The Hearsay-II speech-understanding system: Integrating knowledge to resolve uncertainty*. Computing Surveys, 12(2):213–253, June 1980. See also www.thecepblog.com/2008/07/20/a-brief-introduction-to-blackboard-architectures/, 20.07.2008.

Mountains of Text. Analyzing Alpine Literature from the AAC

Biber , Hanno

hanno.biber@oeaw.ac.at

Austrian Academy of Sciences, Austria

The AAC-Austrian Academy Corpus has already collected and is in the next phase of its development going to collect a considerable number of alpine texts of various kinds from the time between 1848 and 1989. This poster will present a research initiative where the AAC will integrate alpine texts for corpus linguistic research of such a thematic subcorpus, in which the discourse about mountains, that is determined by historical, sociological, cultural and other factors, can be analyzed from various perspectives. Alpine club yearbooks, literature, journals and other related heritage documents will be prepared for analysis. The AAC Alpine collections will follow to some extent the Swiss sister project “textberg.ch”, but with a somehow different research perspective modelled according to the AAC principles. The AAC is a German language corpus of texts from various linguistic domains with around 500 million tokens, operated by the Institute for Corpus Linguistics and Text Technology at the Austrian Academy of Sciences. The question is how corpus research methods based upon a multidisciplinary combination of corpus linguistic and cultural studies can be applied to gain insights into the textual representations of historical collections, in particular alpine texts. “Quantitative corpus linguistics has proved to be a valuable technique in many domains of philological, sociological and historical research. The digitized and linguistically annotated corpus is therefore an interesting source for studies in many fields and facilitates the investigation of changing patterns of language use, and how these reflect underlying cultural shifts.” (Volk, M. et al.: “Challenges in building a multilingual alpine heritage corpus”, LREC 2010)

The AAC will go beyond a quantitative approach and integrate text studies into its research. The methodology of corpus based text research is determined by corpus linguistic, lexicographic and analytical procedures. The language in significant alpine texts of certain historical periods will be of interest. The historical condition of Austria during the Habsburg monarchy with its cultural and linguistic diversity and the situation at the time of National socialism have to be taken into consideration as historical changes with significant influences on the language.

A special emphasis will be given to the culturally productive relation between the metropolis Vienna and its surrounding alpine regions. The Alpine Regions as cultural topographies are if one considers only the places and place names objects of interest and points of interest worth being investigated by means of a critical analysis. In the texts the place names and the thematic complexes associated with the various places are substantial elements. These places of interest are constituted in the texts as places with a particular language, with particular vernaculars, with particular styles, of typical modes of speaking, of typical cultural settings, with typical protagonists and so on. In this study they will also be treated in comparison with other highly important places, such as Alpine regions outside Austria as well as with Vienna, and so on. In this paper special

emphasis will be given to issues of textual representation in connection with linguistic transformation in a broad sense. Investigations into the use of – to give just one example of the research potential – typical idiomatic expressions, into the properties of multi word units in metaphorical constructions of alpine origin, and into fixed forms in particular can be considered as investigations into the properties of figurative language units in general. The analysis of such constructions can be modelled with the help of new methods making use of digitally available sources and techniques. The exemplary online editions of „AAC-FACKEL“ and “Brenner online” of the Innsbruck based journal offer fully searchable databases of the journals with various indexes and search tools in a web interface, where all pages of the original are available as fully searchable digital texts and as facsimile images, all equipped with various indexes and search tools based upon corpus linguistic research parameters.

References

AAC-Austrian Academy Corpus: AAC- FACKEL.

Online Version: "Die Fackel. Herausgeber: Karl Kraus, Wien 1899-1936". AAC Digital Edition No 1, (Editors-in-chief: Hanno Biber, Evelyn Breiteneder, Heinrich Kabas, Karlheinz Mörtl; Graphic Design: Anne Burdick) www.aac.ac.at/fackel .

AAC-Austrian Academy Corpus und Brenner- Archiv:

Brenner Online. Online Version: *Der Brenner*. Herausgeber: Ludwig Ficker, Innsbruck 1910-1954". AAC Digital Edition No 2, (Editors-in-chief: Hanno Biber, Evelyn Breiteneder, Heinrich Kabas, Karlheinz Mörtl; Graphic Design: Anne Burdick) www.aac.ac.at/brenner .

Biber, H. (2009). Hundreds of Examples of Figurative Language from the AAC-Austrian Academy Corpus. In Corpus-Based Approaches to Figurative Language. A Corpus Linguistics 2009 Colloquium, pages 13-20, University of Birmingham, Cognitive Science Research Papers 09-01.

Dobrovols'kij D., Piirainen E. (2005). *Figurative Language: Cross-cultural and Cross-linguistic Perspectives (Current Research in Semantics*, vol. 13). Elsevier, Amsterdam.

Volk, M. et al. (2010): “Challenges in building a multilingual alpine heritage corpus”, LREC 2010

Volk, M. et al. (2009): “Classifying Named Entities in an Alpine Heritage Corpus”. *Künstliche Intelligenz*, 4/2009, S. 40–43.

Exploring Qualitative Data for Secondary Analysis: Challenges, Methods, and Technologies

Bischoff, Kerstin

bischoff@l3s.de

L3S Research Center

Niederée, Claudia

niederee@l3s.de

L3S Research Center

Tran, Nam Khanh

ntran@l3s.de

L3S Research Center

Zerr, Sergej

zerr@l3s.de

L3S Research Center

Birke, Peter

peter.birke@sofi.uni-goettingen.de

Soziologisches Forschungsinstitut Göttingen

Brückweh, Kerstin

Universität Trier, Neuere und Neueste Geschichte

Wiede, Wiebke

wiede@uni-trier.de

Introduction

A remarkable body of data has been collected in the social sciences by interviewing people or observing peoples' interaction in a variety of situations – qualitative data sources very valuable for contemporary research. Notable projects promoting the re-use of qualitative data are ESDS Qualidata¹ (now UK Data Service) or Bequali². Here, we discuss important challenges in re-using qualitative data for secondary analysis and present first ideas on how to overcome them. This includes exploiting state-of-the-art IT-methods from the fields of Information Retrieval and Data Mining – adapting and integrating them for the digital humanities – as well as methodological considerations based on interdisciplinary work, in our case between computer scientists, historians, and social scientists in the project "Gute Arbeit"³.

Challenges in secondary analysis of qualitative data

We focus on challenges on three main levels: a) making qualitative data **accessible** for secondary analysis, b) making relevant material **findable**, and c) making it **understandable**, i.e., ensuring adequate interpretation.

Accessibility

Efforts towards secondary analysis of qualitative data often have to struggle with researchers' reluctance to make their data – their asset – (digitally) available. One crucial issue is warranting anonymity. The problem is exacerbated with qualitative data since answers are rather uncontrolled and unstructured, making it possible to combine information from various places in an interview, e.g., potentially giving information on an employee's (rather unique) background. There is an inherent conflict: While the data owners may tend to prefer protecting their clients, other researchers will argue for having complete information on interview content and context.

Findability

The ability to select the right primary material is an important precondition for re-analysis. For this, tools for exploring and searching relevant studies, cases/samples, and documents are needed that allow defining various criteria and notions of interesting or "similar" data. Furthermore, when reading and analysing the selected (long) interviews we envision enhanced analysis support, e.g. for re-using and sharing codes and annotations or for within document navigation to snippets of interest.

Interpretability

Understanding context is crucial to correctly interpret utterances of interviewees. Lack of context knowledge ("Not having been there") is usually stated as one of the major concerns regarding the feasibility of secondary analysis⁴. Furthermore, some qualitative approaches consider interactions between researchers and interviewees as essential for interpretation⁵. While for some studies e.g., ethnological field studies, (contextual) data may not be sharable at all, for semi-structured interviews the process of data gathering can be made more transparent⁶. Moreover, when working with data from earlier time periods questions of (the comparability of) socio-cultural macro-context are raised⁷.

Technologies for digitally enhanced secondary analysis of qualitative data

In current practice, qualitative researchers mainly rely on qualitative data analysis tools like ATLAS.ti or MaxQDA or on (quantitative) dictionary-based content analysis tools, e.g., General Inquirer or Diction. Here, we discuss how secondary analysis of qualitative data can benefit from more sophisticated techniques from text mining and natural language processing – especially when systematically combining them to reveal novel usages.

Named Entity Recognition

Automatically identifying persons, organizations, and locations, i.e., so called named entity recognition (NER), is a standard task in natural language processing with tools publicly available, e.g. Stanford Named Entity Recognizer⁸. In secondary analysis, NER can be used for improving search (e.g., faceted search) and contextualization. While non-disclosure agreements and adequate access rights will be the cornerstone of an anonymization strategy, NER can also assist in the anonymization task by finding persons or organizations talked about or by highlighting location names, which possibly provide additional hints to who was interviewed. Identified named entities can be systematically substituted by pseudonyms – storing the mapping safely on a remote place.

Sentiment Analysis

Opinion mining (sentiment analysis) techniques could support the secondary researcher in finding opinionated material, e.g., passages with positive or negative points of view on a particular subject. For example, our project "Gute Arbeit" is interested in how peoples' concepts of "good" work evolved over the last decades. Besides, such techniques may support judging the sensitivity of material, e.g. insults. Direct application of – often vocabulary-based – state-of-the-art sentiment analysis tools (e.g.⁹) to qualitative data is usually not feasible. There are peculiarities regarding the detection of subjective expressions and opinion targets, context dependency, indirect opinions, and ordering or omittance effects. For example, in face-to-face interviews subtle sentiment expressions are common. We are researching how to 'train' machine-learning approaches to better cope with qualitative data.

Topic Modeling

Topic modeling with its prominent representative Latent Dirichlet allocation (LDA)¹⁰ is a statistical technique for identifying the topical structure of large textual corpora. Application to limited size qualitative corpora may require gathering additional training data^(11, 12). Topic modeling techniques can highlight themes - possibly going beyond themes asked for - discussed in (long) qualitative documents. Concept Maps or co-occurrence matrixes are related ideas. For a quick overview, interview contents can be visualized by means of representative topics. For example, topics extracted from a collection of studies show commonalities while comparing topics of individual studies sheds light on specifics. Similarly, Janasik et al.¹³ argue that such text mining procedures can aid both data-driven, inductive research by finding emergent concepts as well as theory-driven, deductive research by checking the adequacy and applicability of defined schemes.

Context Enrichment

There are different kinds and levels of context of the interview, e.g., conversational, situational, regarding the research project, or institutional/cultural¹⁴. While most of these context variables need to be documented by the primary

researcher, IT tools can substantially aid capturing the socio-cultural (macro-)context present at the time of data collection. Using external knowledge bases, e.g. Wikipedia or news corpora, primary data can be automatically annotated and linked with background information (e.g. ¹⁵, ¹⁶, ¹⁷). Changes in socio-cultural context may also be better traceable by topic or word clusters with their evolution tracked over time ¹⁸.

Intelligent Search and Visualization

For fast access to an archive of unknown qualitative studies, intelligent search procedures and advanced visualizations for supporting exploration are crucial: Term clouds, Topics maps, and timelines, e.g. for word (cluster) evolutions. Faceted search is a standard in many web applications allowing to browse data or to filter query results based on facets. For a qualitative data archive such facets can be classical metadata like project or study, year, or author, but also advanced information extracted automatically, like entities or topics talked about. All of these may help the secondary researcher to better define his notion of interesting or "similar" material.

While quite many projects make use of one of the aforementioned techniques, novel usage scenarios result from systematically chaining or plugging in the various components. Of course, to aid digital humanities researchers IT tools need to adhere to best practices in interface and interaction design (usability principles like learnability, robustness). More importantly, scepticism regarding the utility and validity of employing IT techniques in humanities research as well as potential misperceptions of 'hostile takeover' attempts have to be addressed.

First experiences

The technologies discussed hold a lot of promise for supporting secondary analysis, but it is important to carefully fit the way they are offered with the work practices and expectations for secondary analysis. Due to the close collaboration across disciplines as well as the work on concrete secondary analysis tasks, our project "Gute Arbeit" provides a good hands-on opportunity for such user-driven technology adaption. We conducted two group discussions where an early prototype realizing topic modelling via Mallet ¹⁹ was shown as a stimulus to each three humanities researchers. Despite some limitations in perceived quality, in sum our researchers stated an added value regarding data access and exploration – though considerably less for those very familiar with the data. In general, the need for iterative interaction, flexibility, and personalization were put forward by both groups. For example, instead of aiming at automatic topic labelling users want to maintain their own topic labels and as well define relationships between topics or group topics into clusters. Thresholds for probability-based techniques like topic modeling should be adjustable to allow trading off completeness versus specificity. By enabling ranking based on relative (cumulative) topic coverage one can easily focus on the most relevant subset of documents that cover most of the topic in the corpus. Contrasting different subsets of documents matching criteria like study, profession, or time period regarding prevalent topics was mentioned as an interesting further development. Especially for the historian, the time dimension was important. Language and topic evolution could be visualized.

Our experiments showed the need to select and adapt text mining tools carefully - here tailoring the technology to the needs of secondary analysis. The lessons learned through our interdisciplinary, collaborative, agile approach to tool development highlight the methodological strengths of the rapid prototyping process: researchers get to know and trust the techniques better as these early hands-on sessions demonstrate potentials as well as necessary refinements. While it is hard to know your requirements for novel digital research tools before seeing them in action, iteratively providing (imperfect) evolutionary prototypes seems a useful methodology for establishing a common ground for the Digital Humanities.

References

1. www.esds.ac.uk/qualidata/about/introduction.asp
2. www.bequali.fr/
3. www.sofi-goettingen.de/index.php?id=1086
4. **Corti, L., Witzel, A., and Bishop, L.** (2005). *On the Potentials and Problems of Secondary Analysis*. An Introduction to the FQS Special Issue on Secondary Analysis of Qualitative Data. Forum Qualitative Sozialforschung / Forum: Qualitative Social Research, 6(1).
5. **Gillies, V. and Edwards, R.** (2005). *Secondary Analysis in Exploring Family and Social Change: Addressing the Issue of Context*. Forum Qualitative Sozialforschung / Forum: Qualitative Social Research, 6(1).
6. **Irwin, S.** (2013). *Qualitative secondary data analysis: Ethics, epistemology and context*. Progress in Development Studies, 13(4):295-306.
7. **Gillies, V. and Edwards, R.** (2005). *Secondary Analysis in Exploring Family and Social Change: Addressing the Issue of Context*. Forum Qualitative Sozialforschung / Forum: Qualitative Social Research, 6(1).
8. **Finkel, J. R., Grenager, T., and Manning, C.** (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, University of Michigan, USA, June 2005.
9. **Pang, B. and Lee, L.** (2008). *Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval*, 2(1-2):1-135.
10. **Blei, D. M., Ng, A. Y., and Jordan, M. I.** (2003). *Latent dirichlet allocation*. *The Journal of Machine Learning Research*, 3:993-1022.
11. **Zhu, X., He, X., Munteanu, C., and Penn, G.** (2008). *Using latent Dirichlet allocation to incorporate domain knowledge for topic transition detection*. Proceedings of the 9th Annual Conference of the International Speech Communication Association, INTERSPEECH, Brisbane, Australia, September 2008:2443-2445.
12. **Tran, N. K., Zerr, S., Bischoff, K., Niederée, C., and Krestel, R.** (2013). *Topic Cropping: Leveraging Latent Topics for the Analysis of Small Corpora*. *Research and Advanced Technology for Digital Libraries - Proceedings of the International Conference on Theory and Practice of Digital Libraries*, TPDL, Valetta, Malta, September 2013, Springer LNCS, pp. 297-308.
13. **Janasik, N., Honkela, T. and Bruun, H.** (2009). *Text mining in qualitative research: Application of an unsupervised learning method*. *Organizational Research Methods*, 12(3):436-460.
14. **Bishop, L.** (2006). *A Proposal for Archiving Context for Secondary Analysis*. *Methodological Innovations Online*, 1(2):10-20.
15. **Mihalcea, R. and Csoma, A.** (2007). *Wikify!: linking documents to encyclopedic knowledge*. Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, CIKM, Lisbon, Portugal, 2007, ACM, pp. 233-242.
16. **Milne, D. and Witten, I. H.** (2008). *Learning to link with Wikipedia*. Proceedings of the 17th ACM conference on Information and knowledge management, CIKM, Napa Valley, CA, USA, October 2008, ACM, pp. 509-518.
17. **He, J., de Rijke, M., Sevenster, M., van Ommering, R., and Qian, Y.** (2011). *Generating links to background knowledge: a case study using narrative radiology reports*. Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM, Glasgow, Scotland, UK, 2011, ACM, pp. 1867-1876.
18. **Wang, X. and McCallum, A.** (2006). *Topics over time: a non-markov continuous-time model of topical trends*. Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD, Philadelphia, PA, USA, August 2006, ACM, pp. 424-433.
19. **McCallum, A. K.** (2002). *MALLET: A Machine Learning for Language Toolkit*, mallet.cs.umass.edu.

SNAP:DRGN - Standards for Networking Ancient Prosopographies: Data and Relations in Greco-roman Names

Bodard, Gabriel

gabriel.bodard@kcl.ac.uk
King's College London, UK

Depauw, Mark

King's College London, UK

Rahtz, Sebastian

King's College London, UK

The **Standards for Networking Ancient Prosopographies** (hereafter **SNAP**) project aims to address the problem of linking together large collections of material (datasets) containing information about persons, names and person-like entities managed in heterogeneous systems and formats. In doing so we must address a number of complex issues:

- How do we know whether a person in one large dataset is the same as a person with the same name in another dataset with a different format and metadata?
- If person A in one database or network is asserted as being the same as person B in another network, do all the statements asserted about person A in the first database also apply to person B, and if so what status do these assertions have in the new context?
- How do we record and integrate the provenance descriptions of both the data and the assertions to create an effective trust network through which we can assess the reliability of a given statement?
- How do systems cope with very large amounts of data and how do we visualize the amount of information available?
- How do we best manage the shift from human-assertions to computer-guided assertions as the dataset moves from human-manageable to big data systems without losing academic credibility in the processed results?
- What academic statements need to be supported to allow the migration of prosopographic and onomastic data silos to interconnected and open data networks?
- What lessons can we learn from working with person-like entity networks, when freed from the privacy and related ethical considerations that would affect modern social networks

These are questions of great interest, and difficulty, for Web scientists studying the contemporary world, given the proliferation of social networks, of different accounts, usernames, handles and URLs associated with the users of the Web, and of databases containing personal information about celebrities, authors, creators and historical figures. **SNAP** will address the more limited, and tractable, problem of datasets containing information about historical names and persons, whether these be onomastica, prosopographies, biographical collections, or merely indices of names from a digital text edition. Specifically, **SNAP** will take the coherent, geographically and chronologically constrained field of prosopography in the Classical world, in which there is established academic activity, as a delimited and realistically sized pilot project to explore solutions to this problem of networking person datasets. Using a selection of the most significant datasets in this sub-domain, we shall address some of the issues of integrating heterogeneous data at the identity level through access to very large but manageable numbers of entities, with the collaboration of the scholars and scientists responsible for the projects involved.

Initially working with a consortium of three collections of person/name data, **SNAP** will expand through additional partners and datasets from the academic, heritage, and commercial sectors in the early stages of the project. Assistance for the inclusion of new datasets will be provided by the project team, as part of the mechanism to test and refine the proposed data models, ontologies and schemas.

These models build on existing work in the field (see Research Context) as well as being designed specifically for this data type, for compatibility and suitability across differently formatted collections. The model's fitness of purpose will be further tested through building distribution and visualization tools, with Web services (see WP3) on top of them; experimenting with the generation of data through computational techniques to leverage the new connections made for new research questions; and expanding our existing, combined dataset with new entities, identifications and connections drawn from inscriptions.

The combined datasets that the project will initially be focusing on contain over 400,000 identified person-like entities and a similar number of attestations, name-entities and annotations records. This number will rapidly expand with the creation of the entities and relationships required by the project model and through the addition of new material and datasets. We envision that the project will be working with over one million entities (persons, names, person-like entities) and their associated statements of relationship, reference and description. These records are not only big data in their own terms but represent a significant portion of the personal data available for the domain and period under investigation. As such this project has the capacity to transform the way that we understand and interact with the data and the related scholarship in the area.

This project will be a proof of concept work; a much larger project will be required to integrate the world of classical prosopography comprehensively, to expand further connections with historical person data from other periods and places, and to test the data models proposed against the more ambitious world of Linked Data as a whole. The project includes a significant time commitment to disseminate and publish the results, both in terms of the ontologies and schemata, guidelines for new projects wanting to participate, and historical articles on information gleaned from this research.

Probing Digital Scholarly Curation through the Dynamic Table of Contexts

Brown, Susan

University of Guelph

Adelaar, Nadine

University of Alberta

Dobson, Teresa

University of British Columbia

Knechtel, Ruth

University of Alberta

MacDonald, Andrew

McMaster University

Nelson, Brent

University of Saskatchewan

Peña, Ernesto

University of British Columbia

Radzikowska, Milena

Mount Royal University

Roeder, Geoff G.

University of British Columbia

Ruecker, Stan

IIT Institute of Design

Sinclair, Stéfan

McGill University

Windsor, Jennifer

University of Alberta

INKE Research Group,

In this paper, we theorize about the role of the curator – or perhaps it would be more accurate to say “custodian” or even “collection designer” – in preparing electronic texts for use in an online digital environment. This scholarly role is becoming increasingly important to new forms of knowledge dissemination as a result of the growth in aggregating or mashing up existing digital content. The goal of these aggregations is to add value for particular purposes, as seen in initiatives such as the *Journal of Digital Humanities*, which aims to collect already published materials into quarterly thematic collections, and the related website *Digital Humanities Now*, which curates weekly the feeds of other digital humanities websites (*Digital Humanities Now*).

By analogy with the definition of a curator as “The officer in charge of a museum, gallery of art, library, or the like; a keeper, custodian” (“curator,” def. n. 6), the digital scholarly curator performs a similar role with respect to the circulation of digital content, though with a number of significant differences. Gallery or museum curators, for instance, have their own scholarly and professional training and preparation, often with respect to proper handling and display of fine art and other valuable physical objects. There are those who feel that at least some curators should be considered artists themselves (e.g. Venzlavov). There is also widespread acknowledgment that curators who work with digital materials require a different set of competencies. Melody Madrid, for example, reports a study that resulted in a list of 20, divided into the categories operational and managerial. Digital scholarly curation also harnesses social media technologies (e.g. crowdsourcing, folksonomies) to encourage a new level of user participation in the management and preservation of digital content (Poole). Not quite an editor, but acting as a mediator between the producers of these contents and its audience through such activities as selecting and reframing them, the digital scholarly curator is in a rather unique position. Our discussion considers this role in relation to the Dynamic Table of Contexts (DToC) interface, a generalized tool for the dissemination of digital text that enables the designer of the collection to mediate specifically between the XML encoding of the text and the affordances that such encoding provides in the reading interface (Brown et al.; Dobson et al.).

The Dynamic Table of Contexts (Fig 1) is a joint initiative between the Interface Design research team of the Implementing New Knowledge Environments (INKE) project, the Canadian Writing Research Collaboratory (CWRC), the University of Alberta Press, and the Voyant Tools project. The DToC currently leverages four principal components of a digital text: the actual text of the document, a table of contents, an index, and XML encoding of the document. The goal of the prototype is to provide an online reading environment where the table of contents provides a conventional overview of a book while at the same time incorporating the index terms and XML tags and the text to which they point (Ruecker et al.). The terms and tags can be selected and deselected, providing interactivity between these three means for accessing and navigating the content of the text or collection.

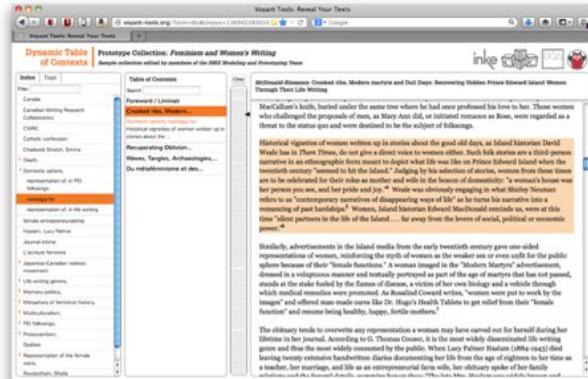


Fig. 1: Dynamic Table of Contexts Interface

The ability to leverage XML markup as part of the navigational interface is a distinguishing feature of the DToC. Customization of that affordance is enabled by what we call the DToC’s “curator mode,” which is distinct from the reading mode in that it allows technically adept superusers to create customized tag lists to serve as navigational aids alongside the index terms, as well as to determine the organization of the table of contents (login is not required, the curated view is expressed through a unique URL). In a print edition, and in particular for anthologies, it is necessary for the editor to decide which of the various alternatives will be used to organize a particular table of contents. For example, a collection of essays might be organized

- alphabetically by title
- alphabetically by author's last name
- chronologically
- by theme
- or by some other principle, in order to ensure a certain kind of development or coherence from beginning to end.

A collection of poems, for example, might add organization alphabetically by the first line of the poem, for cases where the poems do not have a title, and other arrangements are also possible, for instance by geographic location, language, or genre. In the case of an instructor preparing a course pack, the arrangement would naturally correspond to the sequence in which the materials will be used in the class. In the history of print, the possibilities for multiple representation of contents were limited.

In addition to the selection and organization of the contents themselves, curators of DToC collections need to make choices with respect to how the encoding works in the interface. Given the more generic and multi-purpose nature of XML encoding, particularly its use to structure a text, curators need to select which tags the DToC interface will display to the reader, and what user-friendly labels to use for those tags. For many purposes (although not all), the structural encoding can be set aside in favor of semantic encoding (when present). For choices among these tags, there is probably some golden mean that's most appropriate for generic use; the primary benefit of the Dynamic Table of Contexts is to allow variants for more specific uses. Tags that have been rarely used may already be covered by the index, or may be too insignificant to take room on the list. On the other hand, in cases where the tags have been used heavily enough, it may not be useful to include them since it would result in too great a density of hits.

There are also large differences between what kind of curation of tags is required – depending on whether a schema or tagset has instead been applied throughout a collection, such as in the Brown Women Writers Project's use of the Text Encoding Initiative – as opposed to highly customized versions of the TEI adapted to the needs of a particular text. Finally, the curator is also enabled to label how tags will appear in the interface, so that readers are not asked to decipher cryptic forms such as <biblStruct>, but are instead presented with labels for such tags that are meaningful in the context for which the curated text or collection is being prepared. This functionality is useful in cases where there is a nuanced difference between two similar tags that needs to be conveyed to the readers (Fig 2).

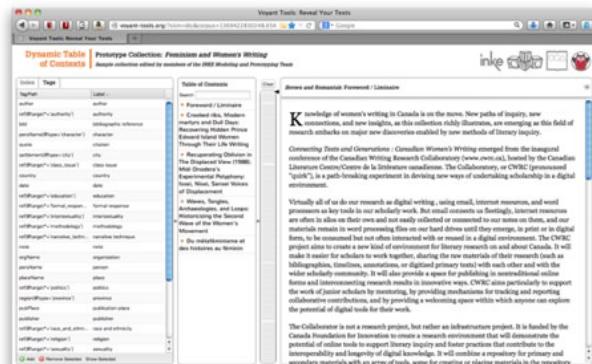


Fig. 2: Dynamic Table of Contents Curator Interface

What this means in terms of the training and qualifications of the DToC curator is that the person needs to hypothesize how the target readers will be dealing with the material, and to use the curatorial functions to customize the view of the text in the DToC interface to meet the anticipated use case(s). In the case of class instructors, to a certain extent the job will be simplified since the course pack has been chosen purposively for the class. In other situations, there may be multiple and possibly conflicting anticipated use cases. The curator also needs to be comfortable enough with XML not only to choose appropriate tags and rename them, but also, if needed, to specify XPath queries to locations in the document that cannot be identified through tag names alone. There will be considerable variation in XML expertise amongst curators, and a previous user study of ours on an earlier version of DToC indicated that a closer relationship to the encoding correlated with more positive experiences of the DToC interface (Dobson et al.).

The paper will frame our understanding of digital scholarly curation in relation to more traditional, historical understandings of curation and representation of contents and will demonstrate the curator mode in the DToC interface. Our previous user studies of the interface found both considerable confusion on the part of users with respect to the role of the encoding or tagging in the DToC (Brown et al.), and, among those who understood it, considerable emphasis on the importance of the XML markup and its ability to shape the reader experience in the interface. As one user said in mousing over the XML Markup pane: "Well, this seems to me the most relevant section, so whoever puts that together is pretty much the wizard in this Oz" (Dobson et al.). Our discussion will incorporate results of the next user study we are conducting on the DToC, with a stress on the curator mode, that will probe these findings which go to the heart of the DToC's affordances. Our aim will be to gain a fuller understanding of the ways in which users with a range of technical knowledge understand the role of markup in the DToC interface, and their understanding as both readers and curators of the curatorial role. This study will inform our understanding not only of the ways in which reading environments such as the DToC can effectively leverage XML encoding, but more generally of the ways in which the idea of curation is rapidly evolving within the online scholarly environment.

References

- Brown, Susan, Brent Nelson, Stan Ruecker, Stéfan Sinclair, Nadine Adelaar, Ruth Knechtel, Jennifer Windsor and the INKE Research Group.** "Text Encoding, the Index, and the Dynamic Table of Contexts." Paper presented at the annual Digital Humanities conference (DH2013), Lincoln, Nebraska. July 16-19, 2013
- "Curator." Def. n. 5. *The Oxford English Dictionary*. 2013. OED Online. Web. 1 Nov. 2013.
- Digital Humanities Now**. <http://digitalhumanitiesnow.org/>
- Dobson, Teresa, Brooke Heller, Stan Ruecker, Milena Radzikowska, Mark Bieber, Susan Brown, and the INKE Research Group.** "The Dynamic Table of Contexts: User Experience and Future Directions." Paper presented in the panel "Designing Interactive Reading Environments for the Online Scholarly Edition" at the DH2012 conference, Hamburg, Germany. July 16-20, 2012.
- Kholeif, O.** *The Curator's New Medium. Art Monthly*. February 2013, (363):9-12. Ipswich, MA.
- Little, G.** *Thinking Like Curators*. Journal of Academic Librarianship, 2013, 39(2), 123-125. doi:10.1016/j.acalib.2013.01.003
- Madrid, Melody M.** *A study of digital curator competences: A survey of experts*. The International Information & Library Review, Volume 45, Issues 3-4, December 2013, pp 149-156.
- Poole, Alex H.** "Now is the Future? The Urgency of Digital Curation in the Digital Humanities." *Digital Humanities Quarterly* 7:2 (2013).
- Ruecker, Stan and the INKE Research Group.** "Introducing the Dynamic Table of Contexts for Scholarly Editions." Paper

presented at the Modern Language Association (MLA) Conference. Los Angeles, CA. Jan 6-8, 2011.

Ruecker, Stan, Susan Brown, Milena Radzikowska, Stéfan Sinclair, Thomas M. Nelson, Patricia Clements, Isobel Grundy, Sharon Balasz, and Jeff Antoniuk. "The Table of Contexts: A Dynamic Browsing Tool for Digitally Encoded Texts." In *The Charm of a List: From the Sumerians to Computerised Data Processing*. Ed. Lucie Dolezalova. Cambridge: Cambridge Scholars Publishing, 2009. pp. 177-187.

Ventzislavov, R. (2014), *Idle Arts: Reconsidering the Curator*. *The Journal of Aesthetics and Art Criticism*, 72: 83–93. doi: 10.1111/jaac.12058

The CWRC-Writer Bridge: from Coder to Writer, XML to RDF, DH to Mainstream

Brown, Susan

Department of English and Film Studies, University of Alberta, Canada; susan.brown@ualberta.ca
School of English and Theatre Studies, University of Guelph, Canada

Brundin, Michael

brundin@ualberta.ca
Canadian Writing Research Collaboratory, University of Alberta, Canada

Chartrand, James

jcchartrand@gmail.com
Open Sky Solutions, Hamilton, Ontario, Canada

Knechtel, Ruth

rknchtle@ualberta.ca
Canadian Writing Research Collaboratory, University of Alberta

MacDonald, Andrew

andrew_james_macdonald@yahoo.com
Open Sky Solutions, Hamilton, Ontario, Canada

Rockwell, Geoffrey

grockwel@ualberta.ca
Humanities Computing, University of Alberta, Canada

Sellmer, Megan

sellmer@ualberta.ca
Humanities Computing, University of Alberta, Canada

Matt Kirschenbaum argues that "the story of writing in the digital age is every bit as messy as the ink-stained rags that would have littered Gutenberg's print shop or the hot molten lead of the Linotype machine" (Schuessler) but the digital humanities community has paid surprisingly little attention to the interfaces that have impacted digital writing, particularly with respect to the tools that enable text encoding or markup. The field has been founded on an understanding of the value of text markup for encoding, preservation, and interoperability. DH projects employ markup consistently in major text-oriented DH applications ranging from linguistics through the production of critical editions to the production of born-digital scholarship and scholarly journals. Yet DH scholars tend to use alternative tools such as mainstream word processors for extensive scholarly writing. Even for submissions to our annual conference, we do not require our community to submit in the form of TEI-encoded texts, even though months are invested annually preparing the book of abstracts for submission. By contrast, submissions to most conferences in the scientific community are routinely produced by the submitting scholars in LaTeX, minimizing the production costs of publishing the submissions (Gauduel 2006). Furthermore, there has been little extension of awareness into the broader scholarly community of the value of XML in general and the TEI in particular, and consequently little uptake beyond the DH community itself.

This is not because humanities scholars cannot use markup. The pervasive use of wikis, blogging software, and social networking sites that allow limited markup indicate that the principles of markup, at least as they relate to on-screen rendering in HTML, are easily acquired. However, such

contexts are conducive neither to an understanding of the nuances and procedural affordances of complex markup nor of how it can structure an entire document; the markup is simply an isolated means to a presentational end. So there is still a considerable gulf between the knowledge base of the mainstream scholarly community and that of the DH community. This situation is exacerbated by the fact that use of XML still demands that users acquire and install editing software that is divorced from the usual contexts of scholarly production and because such packages require a degree of setup that is daunting. Furthermore, the interfaces of XML editing programs tend to be far removed from those with which mainstream scholars are familiar. That of the oXygen XML editing package, for instance, is closer to the look and feel of a Java editor, a programming environment, than it is to a WYSIWYG writing environment. Such applications are not conducive to use by humanities scholars, whose main activities are writing and editing. Members of the DH community have vocally opposed the concept of WYSIWYG XML editors on the grounds that they would undermine a writer's understanding of the function of markup, although this is not a technical consideration: oXygen's libraries have been designed to support the production of interfaces that can allow "non technical users to encode information in XML without actually knowing anything about the underlying XML format" (Bina 2013). There thus exists a tension between the requirements of technically adept super-users, who want power from their tools, and the basic usability of editing interfaces with respect to writing, one that has been there from the early days of markup editors (Karney 1995).

The CWRC-Writer is the centerpiece of an online research environment, the Canadian Writing Research Collaboratory (CWRC), designed to support research on the study of writing in and about Canada. It aims at a user base of mainstream literary scholars, and recognizes the gap in knowledge between that community and digital humanists with respect to best practices for digital resource production. The CWRC-Writer is meant to work with the Collaboratory's other online tools to help bridge that gap by facilitating the production of born-digital scholarship and primary text editions in XML. CWRC will provide an online repository to house digital objects for members of its research community. These will range from bibliographical records, granular chronology entries, profiles of authors and other historical persons, prosopographic data and other entity records, and transcriptions of primary texts, as well as other document types, images, audio and video, although the emphasis remains on writing. CWRC encourages collaboration, whether or not scholars work together formally on a specific project. The CWRC infrastructure supports sharing, reuse, and continual enhancement of scholarly materials, as well as open-access dissemination. Thus one of the most common use cases we anticipate is that someone who is using CWRC for their research sees an error or an opportunity for enhancement within an existing object. We want to make it as easy as possible for that scholar to correct the OCR error, add missing bibliographical information, clarify an ambiguity, or smooth out some infelicitous language. The scholar should not need to download specialized software, but rather move easily from a reading interface to a production interface within the same browsing environment, adding further value to the scholarly resource quickly and easily, in keeping with the realization that we need to overcome siloage between applications (Bradley and Hill 2011).

This is our main use case. Nevertheless, we do have amongst our user community serious textual editors who plan to create digital editions. Our assumption has always been that for heavy-duty markup or transformations one would need to go outside the CWRC-Writer environment to a full-featured XML editor. These advanced users are understandably testing the CWRC-Writer interface from the perspective of their full range of needs. Moreover, since we began publicizing development of the CWRC-Writer, we have received expressions of interest from members of the DH community who want to consider it for use in TEI editing projects, as components of library-based DH tool suites, or for teaching XML. The possible use cases for CWRC-Writer are thus situated along a spectrum ranging from the production of born-digital content or primary text editions from scratch, in which case much of the technical demands of

markup application need to be performed within CWRC-Writer itself, through to the quick fixing of errors or editorial revision of existing documents. This spectrum can be elaborated, so that the poles of these two use cases are aligned with the degree of complexity of the interface and indeed of functionality of the editor itself, and the level of expertise expected of the user.

Interface Complexity	<----->	Interface Simplicity
Expert users Production from scratch		Novice users Quick edits/fixes

Simplicity is also relative and contextual. While the CWRC-Writer does not support advanced XML features, its interface has considerable complexity as a result of its integration with other aspects of the Collaboratory. It is more complicated, for instance, than the newly launched DHwriter designed to facilitate the production of abstracts for the DH conference. Because CWRC aims to support interoperability and discoverability by using Linked Open Data entities for authority control, the interface is complicated by the fact that the editor combines the application of XML markup with the annotation of documents with RDF entities. To make this blended approach as seamless as possible, we have mapped our RDF specifications for named entities onto the equivalent tags within supported XML schemas. Thus users identifying a person's name within a text are simultaneously applying a <persName> tag, if the document is using a TEI schema, as well as creating an RDF Open Annotation object. This increases the challenge of interface design in a number of ways. For instance, there are some basic conceptual confusions with respect to terminology, since there are two types of "tagging" available within the editor, sometimes operating in tandem and sometimes not. Tagging means different things within technical vocabularies and mainstream folksonomic contexts, so it is a balancing act to bridge from popular literacies to the CWRC-Writer environment while also retaining sufficient accuracy of terminology to help develop digital humanities literacies. The resulting confusion has emerged as a theme in our user testing to date and brings home the extent to which the CWRC-Writer emerges from an expanded understanding of annotations and their potential to support new paradigms of interoperative and interactive digital scholarly environments (Bradley 2012; Grassi 2013).

We last reported on the CWRC-Writer editor in a pre-alpha state, since which it has undergone numerous development iterations and substantial user testing. By DH2014 we will have conducted our most extensive user testing on the beta version, with users ranging from novices to DH experts. Our aim is not only to improve the functionality of the system as an editor, but to try to understand how the interface works as a writing and encoding environment for various types of users, including asking respondents to compare the interface experience to other writing environments, both Web- and PC-based.

The poster will thus do the following:

Provide two computers on which people can test the CWRC-Writer;
Summarize the major affordances of the editor and the concepts behind it;
Summarize the results of the user testing.

The testing results will allow us to situate the CWRC-Writer as a tool within the current editing landscape, along the spectrum outlined above, and to evaluate the use cases for which the CWRC-Writer is best suited. The poster will facilitate dialogue regarding the relationship amongst text encoding, Semantic Web technologies, and mainstream scholarly writing processes. The testing results will provide insights into the tensions in interface design between expert vocabularies and best practices, on the one hand, and mainstream vocabularies and scholarly pragmatics on the other. We hope to make a contribution to "ways of seeing" (Kirschenbaum 2004) markup environments and relation to digital scholarly production on the Web.

References

Bina, George (2013). "Customizing a General Purpose XML Editor: oXygen's Authoring Environment." Proceedings of the International Symposium on Native XML User Interfaces, 2013. <http://www.balisage.net/Proceedings/vol11/html/Bina01/BalisageVol11-Bina01.html>

Bradley, John (2012). "Towards a Richer Sense of Digital Annotation: Moving Beyond a 'Media' Orientation of the Annotation of Digital Objects." DHQ 6.2 (2012).

Bradley, John and Timothy Hill (2011). "When WordHoard Met Pliny: Breaking Down of Interaction Silos Between Applications." Digital Humanities 2011, Stanford University, June 19–22, 2011. <http://pliny.cch.kcl.ac.uk/docs/Stanford-Poster.pdf>

Brown, Susan (n.d.). "Scaling Up Collaboration Online: Towards a Collaboratory for Research on Canadian Writing." International Journal of Canadian Studies. Forthcoming.

Canadian Writing Research Collaboratory. www.cwrc.ca. CWRC-Writer. <https://github.com/cwrc/CWRC-Writer> DHwriter. dhwriter.org

Gaudéul, Alexia (2006). "Do Open Source Developers Respond to Competition?: The (La)TeX Case Study." (March 27, 2006). Available at SSRN: <http://ssrn.com/abstract=908946> or <http://dx.doi.org/10.2139/ssrn.908946>

Grassi, Marco, Simone Fonda, and Francesco Piazza (2013). "Pundit: augmenting web contents with semantics." Literary and Linguistic Computing 28.4 (2013).

Karney, James (1995). "Author/Editor." PC Magazine 7 February 1995. 153ff.

Kirschenbaum, Matthew G. (2004). "So the Colors Cover the Wires: Interface, Aesthetics, and Usability." A Companion to Digital Humanities, ed. Susan Schreibman, Ray Siemens, John Unsworth. Oxford: Blackwell. <http://www.digitalhumanities.org/companion/>

Rockwell, Geoffrey, Susan Brown, James Chartrand, and Susan Hesemeier. (2012). "CWRC- Writer: An In-Browser XML Editor." Digital Humanities 2012 Conference, Hamburg, July 16–22. Digital Humanities 2012 Conference Abstracts. Hamburg University.

Schüssler, Jennifer (2011). "The Muses of Insert, Delete and Execute." New York Times, 25 December 2011.

Wuppertal for the project *Arthur Schnitzler: Digitale historisch-kritische Edition (Werke 1905–1931)*. While it will be available as an independent stand-alone version, it forms the basis of the project's virtual edition platform which significantly expands its functionality.

Project and Requirements

Funded by the *Nordrhein-Westfälische Akademie der Wissenschaften* and collaborating with *Deutsches Literaturarchiv Marbach*, our long-time project aims to create a digital critical edition that makes available both printed works and Schnitzler's estate in an up-to-date and philologically dependable form.

Next to the edited text, the focus will be equally on the materiality and the genetic dimension of the work, which involves various methodological challenges. Along with the complexity and sheer size of the underlying material (> 12000 pages to transcribe) it became obvious very soon that the objectives were not to be attained with existing software solutions (being mostly optimized for the reworking of existing transcripts).

Transcribo was therefore developed from scratch for the efficient creation of large amounts of differentiated, deeply annotated transcriptions, which is reflected in its entire layout.

Transcribo: Features and Interface

The graphical user interface is centered around the digital facsimile, usually the scanned physical witness. This facsimile is duplicated, so that the original can always be viewed unobstructed, while all processing steps take place on the slightly grayed duplicate. This arrangement accommodates the use of multiple monitors and particularly saves a time-consuming jumping back and forth between the image and editor window. Thus, selections in rectangle or polygon shape can be drawn topographically precisely above the graphic and the transcribed text can be entered directly. This also simplifies, if desired, to proceed in a non-linear way and create the transcripts in any order, for instance to treat graphically related, but spatially separated locations in one step. Furthermore, an OCR (with optional image optimization such as contrast and color adjustment) is integrated for the detection of typescripts, providing a raw transcription for the editors to build upon.

In doing so, it is essential that any recorded unit can be commented and that philologically and genetically relevant phenomena can be annotated in a uniform way. This is achieved by context menus with a wide and project-specific choice of options. So far, this includes different variants of corrections and changes inserted by author or editors, the labeling of unsafe readings or unidentified graphs, the marking of graphical elements or global attributes of witness and writing instruments. This selection can be extended at will and will be adapted to the textual requirements over the entire course of the project.

Also, transcribed units can be combined into sequences, whether to document comprehensive semantic or genetic correlations such as combined changes or, in case of extensive overrides and insertions, to mark the resulting textual order.

Since each facsimile is a separate unit, no cross-page connections can be documented in the stand-alone version of the software, a shortcoming that is remedied by its connecting to a database system.

Transcribo: A Graphical Editor for Transcribing and Annotating Textual Witnesses. Preparing a Historical-Critical Edition of Arthur Schnitzler's Works.

Buedenbender, Stefan

Universität Trier, Germany

Friedrich, Vivien

Bergische Universität Wuppertal, Germany

Burch, Thomas

Universität Trier, Germany

Fink, Kristina

Bergische Universität Wuppertal, Germany

Wolfgang, Lukas

Bergische Universität Wuppertal, Germany

Kathrin, Nühlen

Bergische Universität Wuppertal, Germany

Frank, Queens

Universität Trier, Germany

Joshgun, Sirajzade

Universität Trier, Germany

Transcribo is a graphical editor for transcribing and annotating textual witnesses, which is being developed at *Trier Center for Digital Humanities* and *Bergische Universität*

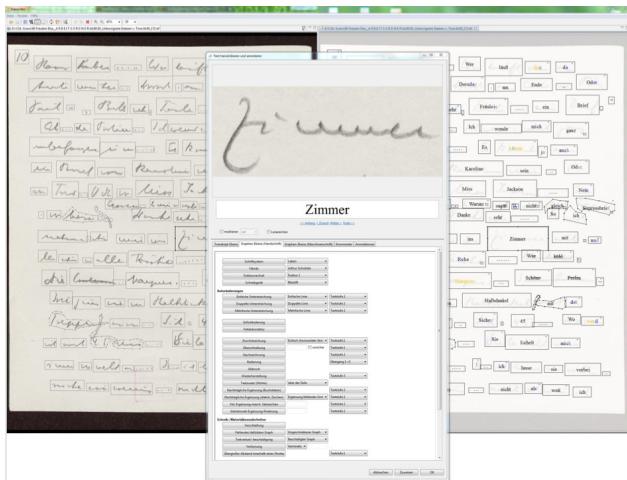


Fig. 1: Transcribing and annotating a manuscript

Transcribo and FuD: Extending the Reach

At this point, however, a basic methodological problem arises: The superposition of a semantic, genetic, and material-documentary perspective creates different levels of description which overlap and intersect. Such configurations are very cumbersome to handle by exclusively XML-based data structures or software solutions. This is one of the reasons why the developers refrained from letting the users work directly in an XML environment. Instead, the program is based on an internal data structure, the complexity of which is attenuated by the graphical interface, while the results can be exported at any time to XML/TEI.

Moreover, it allows for a seamless integration into an existing environment, the Research Network and Database System "FuD"^[i].

Transferring the content into a centralized database system and associated infrastructure further expands the capabilities. The existing options for annotation can thus be extended to sequences of any size, independently of given material or logical units. Furthermore there are already proven tools for the managing of metadata and marking of related motifs. Also, auxiliary functions such as a graphemic search can be implemented: if an editor is uncertain about a specific transcription, all previously recognized graphical equivalents to a given sequence of letters can be displayed. As outlined in our last year's contribution, further program modules are currently being developed which allow the visualization and the comparison of writing stages.

References

- www.arthur-schnitzler.de (will go online on Nov. 15th, 2013). For details, see also: dh2013.unl.edu/abstracts/ab-412.html. Forschungsnetzwerk und Datenbanksystem: fud.uni-trier.de.

Digitalizing the Matsu Festival Celebration: The Study and Application of Value-Added Creative Methods to Taiwan Folk Culture and Art

Chen, Chun-Wen
junbun@mail.cgu.edu.tw
 Chang Gung University

Hsu, Su-Chu
suchu.hsu@gmail.com

Taipei National University of the Arts

Day, Jia-Ming

jimmyday2010@gmail.com

Taipei National University of the Arts

Lin, Cheng-Wei

moriokastr@gmail.com

Taipei National University of the Arts

In August 2009, *Mazu belief and customs* were formally inscribed on UNESCO's Representative List of the Intangible Cultural Heritage of Humanity (2013)¹. Worship of Mazu has long been the folk belief with the largest numbers of believers and temples in Taiwan. It is a deeply rooted faith for Taiwanese. The folk belief has been developed from Taiwan's immigrant society and maritime culture. Mazu is a goddess transformed from an ordinary female who can protect people on the sea from disasters. Worship of Mazu has spread to more than 20 countries around the world, and has had a profound impact on the majority of the faithful. The artist Chin-Hsin Lin has spent 20 years completing the giant 124 m woodcut prints *Celebrating the Mazu Festival*. It records the celebrations of the most important religious deity in Taiwan with rich details. The prints, exhibited in many museums abroad, have received international recognition. The *Celebrating the Mazu Festival* prints are not only artworks but also a part of Taiwan folk culture. Because of the richness of the work, we believe the cultural spirit and content can be expressed through multiple-sensory interactions in digital media.

Based on the longest woodcut print scroll in the world, *Celebrating the Mazu Festival*, one of the most important Taiwan folk cultural artwork, we propose an interdisciplinary project *Digitalizing the Mazu Festival Celebration: The Study and Application of Value-Added Creative Methods to Taiwan Folk Culture and Art*. We use digital methods to create it with value-added applications of digital media, such as animation, web site, interactive installation, and mobile app. The goals of the project are as follows:

1. To complete Taiwan's first digital animation of *Celebrating the Mazu Festival*.
2. To add value to the longest woodcut print scroll in the world by applying interactive motion-sensing technology in installation art.
3. To use the *Celebrating the Mazu Festival* folk culture material to construct Web 2.0 and mobile social media with creative application design.
4. To integrate humanistic art, media technology, and cultural and creative industries, to develop the new concept "interactive & digital" culture in Taiwan, and to transfer the experience to education and industry.

This integrated project consists of three sub-projects: (1) The Analysis and Design of Cultural and Creative Content: content analysis and animation creation; (2) The Study and Development of Interactive Display Technology: development of motion-sensing and mobile technologies; and (3) The Application of the Innovative Display Model: design and construction of the web exhibition and interactive installation. The first sub-project uses the Mazu Festival Celebration theme and create varied artistic forms with animated image technology to present Taiwan folk culture. In the second sub-project, we study and develop motion-capture and related innovative display technologies.

This paper focuses on the third sub-project that applies multiple senses and interaction in digital media to represent the cultural content. Integrating information graphics, navigational design, and a dynamic visual style is the design approach. This sub-project consists of three parts: (1) Web Exhibition Hall: To design the information architecture and navigation system, to design and produce the *Web Exhibition Hall* with 67 prints on the exhibition web; (2) *Web Creative Displays*: To create *Web Creative Displays* on selected web pages with interactive information graphics to create deep understanding of the cultural content; (3) Interactive Installation: To design and build the physical installation, and to integrate animated content, artistic form, and motion-sensing technologies in the "*Celebrating the Mazu Festival*" *Interactive Installation*. The

process and results show the possibility of integration of folk art and digital media.

References

1. UNESCO. (2013). *Mazu belief and customs*. Retrieved January 20, 2013, from www.unesco.org/culture/ich/index.php?lg=en&pg=00011&RL=00227.

Arabic and Greek New Testament manuscripts: Identities and Digital cultures

Clivaz, Claire

University of Lausanne (CH)

Schulthess, Sara

University of Lausanne (CH)

Bouvier, David

University of Lausanne (CH)

Teule, Herman

University of Nijmegen (NL)

This research project is supported by the Swiss National Fund under the form of a 3 years grant for a PhD in Arts and Humanities on New Testament, Ancient Greek and Christian Arabic traditions, as interdisciplinary project (2013-2016). It tries to analyse the global disinterest of the Western research for the Arabic manuscripts of the New Testament (NT) since 1945, and the recent growing-up attention for them among Middle-East Christians and some Western scholars, as well as the interest of certain Muslim circles for the Greek and Arabic NT manuscripts, in particular on websites. This project explores the necessary transformations of a classical philological field to study such a phenomenon. The digital culture is the common factor that allows to Western and non Western scholarship to cross over these ancient objects: the digital support of writing requires to use sociology, philology, history, and epistemology to analyse the hybrid cultural objects that it creates.



Fig. 1: www.sheekh-3arb.net/bible

The first step of the project is to enlighten, according to the approach of the cultural studies, the history of research of the field. The second step is to demonstrate the usefulness of the Arabic manuscripts by doing and studying the list of the Pauline Arabic manuscripts, and to investigate the weight of the language and the culture on the text by editing the First Corinthians in the manuscript Vat. Ar. 13. The third step is to understand the identity polemics on the Christian and Muslim websites studying Greek and Arabic manuscripts of the New Testament. The Digital Humanities epistemology offers here the general background to decipher what happens in a classical philological field of study.



Fig. 2: Codex Sinaiticus (Mk 1.1) on a Salafist Muslim website; <http://www.sheekh-3arb.net/vb/showthread.php?t=2127&page=3> (06.03.2014)

Beyond the printed culture, the digital support of writing leads by itself to reconfigure the boundaries of knowledge: in such a project, not only diverse ancient languages are required (Greek, Arabic, Latin), or classical codicology and philology, but also deep epistemological inquiries, with the help of sociology (network analysis), pedagogy (multiliteracies), and other disciplines. Finally, one assists to the emergence of an interdisciplinary terra incognita: a hybrid cross-cultural scholarship. New concepts and points of view are here to be used and created.

More generally speaking, this project participates to the debate about new models of digital critical edition: are we ready to renounce to a stabilized text, in favour of open-ended texts on websites, allowing to integrate comments, diverse languages and images of manuscripts? The digital edition of Homer is going in the sense of a history of reading, rather than to offer a "definitive" text (<http://www.homermultitext.org>). The example of the present research project underlines that the notion of text is more and more replaced by the "document" in a digital cultural perspective. This project hopes also to demonstrate that we are going from a knowledge based on the general relationship subject-object, to an intersubjective knowledge, developed in networks: for example, how are we going to integrate new forms of academic productions in our analysis, such as blogs and Internet forum?

In this mind, we want to challenge the complexity of our project with a website (www.unil.ch/nt-arabe) offering the possibility of going above the limits of the printed media.

References

- Arbache S.**, *L'Évangile arabe selon saint Luc. Texte du VII^e siècle, copié en 897*. édition et traduction, Bruxelles, Safran, 2012.
- Clivaz C.**, "Homer and the New Testament as 'Multitexts' in the Digital Age?", in SRC 3 (2012/3), 1-15 ; open access : <http://src-online.ca/index.php/src/article/view/97>.
- Clivaz C.**, "Internet Networks and Academic Research: the Example of the New Testament Textual Criticism", in Clivaz C., Gregory A., Hamidovic D., in collaboration with Schulthess S. (eds.), *Digital Humanities in Biblical, Early Jewish and Early Christian Studies* (Scholarly Communication Series 2), Leiden, Brill, 2013.
- Clivaz C., Hamidovic, D.**, "Critical Editions in the Digital Age", in *The Johns Hopkins Guide to Digital Media and Textuality*, Ryan M.-L., Emerson L., and Robertson B. (eds.), forthcoming.
- Griffith S.H.**, *The Bible in Arabic: The Scriptures of the 'People of the Book' in the Language of Islam*, Princeton/Oxford, Princeton University Press, 2013.
- Kashouh H.**, *The Arabic Versions of the Gospels. The Manuscripts and their Families* (Arbeiten zur neutestamentlichen Textforschung 42), Berlin, de Gruyter, 2011.

New London Group. "A Pedagogy of Multiliteracies: Designing Social Futures", in *Harvard Educational Review* 66, no. 1 (1996), pp. 60–92.

Schulthess S., "Die arabischen Handschriften des Neuen Testaments in der zeitgenössischen Forschung: ein Überblick", in *Early Christianity* 3(4) (2013), pp. 518–539

Schulthess S., "Les manuscrits du Nouveau Testament, le monde arabe et le digital. L'émergence d'un discours hybride", in Clivaz C., Meizoz J., Vallotton F., Verheyden J. (eds.), in collaboration with Bertho B. (eds.), *Lire Demain. Des manuscrits antiques à l'ère digitale / Reading Tomorrow. From Ancient Manuscripts to the Digital Era*, PPUR, 2012, pp. 333–344

Thomas, D.R. (ed.), *The Bible in Arab Christianity*, Leiden, Brill, 2007.

Empowering Play, Experimenting with Poems: Disciplinary Values and Visualization Development

Coles, Katharine

katharine.coles@utah.edu

University of Utah, United States of America

Meyer, Miriah

miriah@cs.utah.edu

University of Utah, United States of America

Lein, Julie Gonnering

julie.gonnering@utah.edu

University of Utah, United States of America

McCurdy, Nina

nina@cs.utah.edu

University of Utah, United States of America

At DH2013 we presented our initial approach to poetry visualization, which undertakes to support active, spontaneous, and unique close reading experiences by emphasizing temporality and reader-engagement in exploring over twenty dimensions of poetic sound (Abdul-Rahman et al., 2013a; Abdul-Rahman et al., 2013b). PoemViewer, the software we discussed last year, makes important advances toward that aim; some literary colleagues already proficient with digital tools have noted how they might use that program in their own research and classrooms. But many still wonder, *Why not just read the poems?* Sometimes that question indicates a simple lack of digital literacy or a defensive response to a perceived technological threat. But it also points to ongoing tensions between core values and practices in the humanities (like close reading) and those implicit in the engineering that makes digital tools (like data visualization) possible.

People will always read and write poems without computers. Still, poets like Nick Montfort and Stephanie Strickland compellingly bridge literary composition and computational techniques. And DH research has powerfully aided traditional humanities scholarship in archival work, mapping, etc., that would also have continued, if differently, without computational tools. We too strive to design software that can support existing humanities practices while also suggesting new ones—a dual goal we believe is served by honoring different disciplinary values: how might we in practice value multiplicity and openness as well as exactitude and decision, for instance? Rather than endeavoring to quell such tensions, we argue they should be embraced and productively leveraged through *play*—not only in the capabilities a tool can ultimately deliver, but also in the conception and experimentation that inform each stage of its development.

This poster will share images and anecdotes from the technology probes—what we're calling 'experiments'—the rapid prototyping informing and guiding the design of our new close reading tool, Poemage. We will explain how these first development stages already begin to answer the call from DH scholars like Johanna Drucker and Stephen Ramsay for

digital tools more imbued with and suited to the humanities—without sacrificing scientific meticulousness and rigor. Like PoemViewer, Poemage will invite users to visualize poems of their own choosing, rather than rely on existing visualizations of poems chosen and manually coded by others. By using automated techniques to reveal various nuanced kinds and patterns of sonic ambiguity (or play) within poems, Poemage promises to advance the fields of computer science, literary study, and digital humanities. By embracing play in its development and prioritizing play as a condition of use, it seeks to empower a greater range of scholars and readers actually to engage and benefit from those advancements.

Poems are not reducible to their constituent pieces. Sound affects tone; lineation expands and turns syntax; meaning is multiple, building on ambiguities and multiplicities of diction and rhyme and so forth—and each reader adds to the complexity already inherent in the words and their arrangement every time she reads a poem. Moreover, all of this happens through many dimensions and in multidirectional time, a factor usually overlooked by other poetry visualization software but crucial to our efforts. We believe that poems are not only animated but created and recreated through each act of reading, and therefore that our tools should foreground this act and the relationship that develops between the reader and the poem. Rather than simplifying poetic complexity or stifling ambiguity, then, we accentuate and explore them. This approach to poetry visualization does not save a reader's time, but we do hope our approach enriches time spent. We want users' interaction with our software and visualizations to continually lead them to new ways in and through a given poem. We want them to experience the pleasurable mental focus and the intellectual and emotional expansion that happen through play: the play that inheres in particular poems and the opportunities for play that emerge as the reader explores a poem with our software.

These goals require Poemage to offer an unprecedented degree of *fluidity, precision, and responsiveness*. In the belief that these complementary values are shared by literary and computer science disciplines alike, our team is endeavoring to build them into Poemage by layers and degrees via strategies like rapid prototyping, user-centered design, and agile programming. To that end, we are experimenting with different algorithms for detecting and classifying sonic elements in a poem, as well as quantifying the uncertainty in these computational results, all within the model of play. In our work, much as in close reading, play entails testing the tool in multiple ways through query and response and through continually looking at and through new approaches, even when existing approaches may seem to yield satisfactory answers, with an emphasis as much on experience and surprise as on hypothesis and (inherently uncertain) results.

Our poster will share visualizations and screenshots and discuss what we have learned both from our experiments and from some of the practical challenges we have faced so far. For example, our efforts to capture patterns of morphing sonic clusters (like the e/s/t group recurring in different orders and in different syllables in Louise Bogan's [1968] poem "Night"—"estuaries," "restless," "inlets," "set," "itself," "reflects," "firmament's," and "setting") via data sets and adaptable, customizable queries have raised questions about how much preprocessing is possible and desirable, as well as how much complexity to expose to the user. How do we balance support for open and spontaneous exploration with the computational limitations of computer memory, real-time processing speed, and algorithmic definitiveness? One way is to offer multiple exploratory strategies. Poemage will show readers the set of words in a poem participating in a particular aspect of traditional rhyme, such as a specific vowel recurrence (see Fig. 1). We are also exploring how to show "sets of sets," displaying for instance every distinct assonance pattern (Fig. 2), or the variety of sonic patterns overlapping in a chosen word (Fig. 3). We are also developing a character cluster search, unique to Poemage, which will reveal sets of words containing letter units of varying length in any order (Fig. 4)—but whose algorithm as yet does not stretch to reach past intervening phonemes. Such gaps, though, invite reader participation to extend or reframe computational results.

Interesting computation limitations also expose other kinds of uncertainty and ambiguity for the literary scholar to consider. For example, an early prototype assumed the pronunciation of “wind” in Bogan’s line, “The restless wind of the inlets,” was the verb “to wind.” It therefore linked that long “i” with words like “tide,” “lights, and “behind” elsewhere in Bogan’s poem (see Fig. 5). Results like this might be taken for meaningless, undesirable noise—except that “Night” enacts and meditates on various kinds of movement and so the notion of winding (of certain movements storing potential energies for other, differently expressed movements) is actually highly relevant, even though a reader might not have noticed it without a prompt. Poets are attracted to noise; part of their work is to render noise meaningful and aesthetically pleasing, and we hope in our visualizations not to eliminate but to capture and highlight just these kinds of uncertainty, which enhance rather than detract from the reading experience.

As their different names might suggest, Poemage—even more than PoemViewer—emphasizes the interpretive experience, oriented toward play rather than proof. Relying on but moving beyond the transformational power of PoemViewer’s “magic lens” that can reveal previously undetected but nevertheless unhidden poetic features useful for close reading (including the “surface reading” described by Best and Marcus [2009]), Poemage enhances the imaginative conjuring that can happen in readers’ interaction with poetic language and visualizations. It shares Drucker’s “performative materiality” ethos and works to support the critical and computational practices she outlined in *DHQ*(2013): “In place of transparency and clarity, [...] foreground[ing] ambiguity and uncertainty, unresolvable multiplicities in place of singularities and certainties. Sustained interpretive engagement, not efficient completion of tasks, [is] the desired outcome.” For, as Stephen Ramsay (2011) puts it:

Literary critical interpretation is not just a qualitative matter; it is also an insistently subjective manner of engagement. . . . [C]onclusions are evaluated not in terms of what propositions the data allows, but in terms of the nature and depth of the discussions [—and we might add, experiences—] that result. . . . We are not trying to solve Woolf. We are trying to ensure that discussion of The Waves continues. (8-9, emphasis added; 15)

This is completely consistent with our efforts: we do not wish to "solve" the poem; if anything, we wish to open it to further and deeper questioning, to provide opportunities for continuous *unsolving*. It's what our team is working toward both literally and metaphorically, as we seek to design poetry visualization empowering play.

Captions

Night

Louise Bogan

The cold remote islands
And the blue estuaries
Where what breathes, breathes
The wind of the inlets
And what drinks, drinks
The incoming tide;

Where shell and weed
Wait upon the salt wash of the sea,
And the clear nights of stars
Swing their lights westward
To set behind the land;

Where the pulse clinging to the rocks
Renews itself forever?
Where again on cloudless nights,
The water reflects
The firmament's partial setting,

O remember,
In your carrying dark hours
That more things move
Than blood in the heart

Rhymes	Character Clusters
Identical Rhyme/Rhyme Riche	● * * * *
Perfect Rhyme
Semirhyme	- -
Syllabic Rhyme	* -
Consonant Slant Rhyme	● * * *
Vowel Slant Rhyme	● * *
Pararhyme
Eye Rhyme
Alliteration	● * * *
Assonance	● ● ● EH *
Consonance	● * * *
Forced Rhyme
Syllabic Rhyme v2	* - -
Imperfect Consonant Slant

clear beautiful hairball

Fig. 1: Rhyme set of words in “Night” containing “eh” assonance.

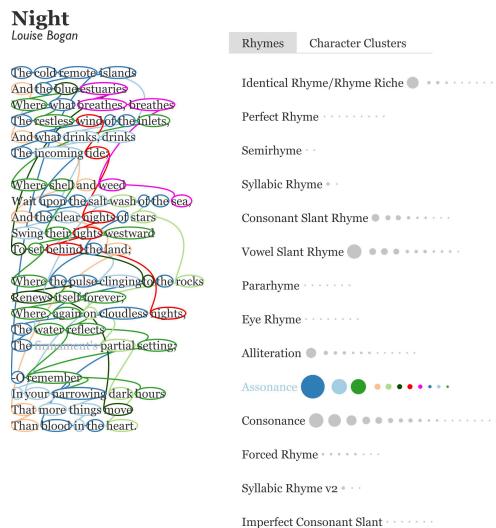
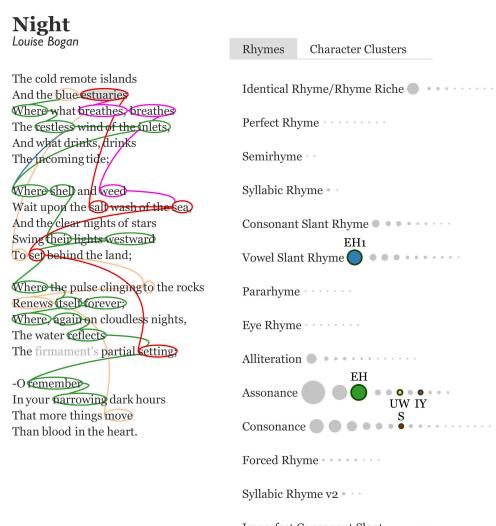


Fig. 2: Set of assonance rhyme sets in “Night”



Journal of Health Politics, Policy and Law, Vol. 35, No. 4, December 2010
DOI 10.1215/03616878-35-4 © 2010 by The University of Chicago

Fig. 3: Set of assonance, consonance, and slant rhyme sets overlapping in the word "estuaries" in Bogdan's poem

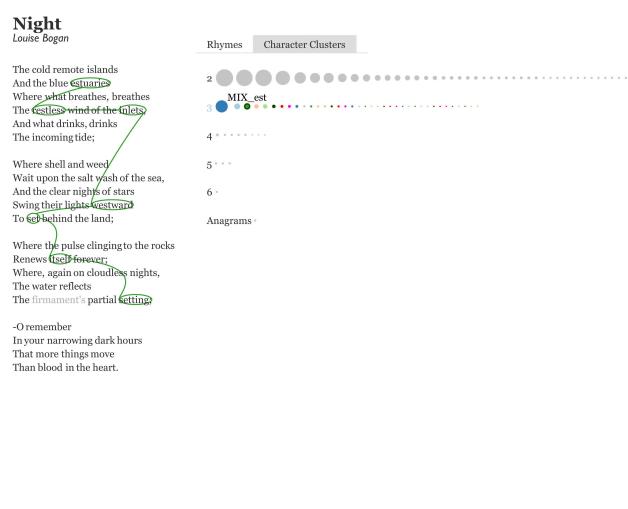


Fig. 4: The E-S-T character cluster set, revealing each instance of these letters appearing together in any order and in any syllable within Bogan's poem. Users can also explore other sets of clusters of varying length, including examples of anagram.

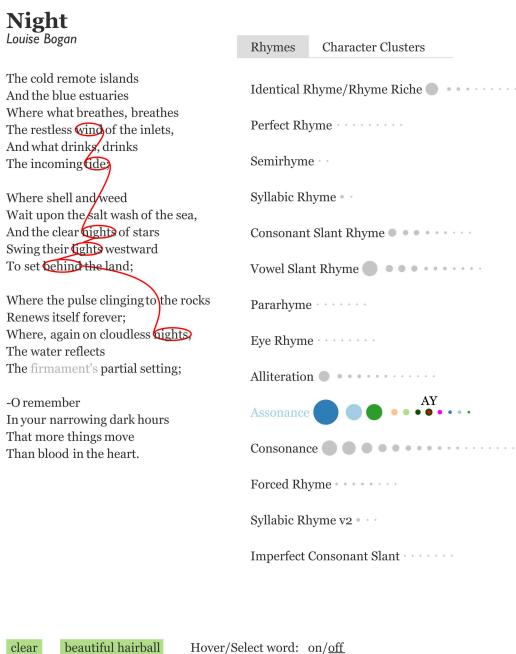


Fig. 5: Rhyme set of words in “Night” containing “ay” assonance.

References

- Abdul-Rahman, A., K. Coles, J. Lein, M. Wynne.** (2013a). Freedom and Flow: A New Approach to Visualizing Poetry. Paper presented at Digital Humanities 2013, University of Nebraska-Lincoln. July 2013.

Abdul-Rahman, A., J. Lein, K. Coles, E. Maguire, M. Meyer, M. Wynne, A.E. Trefethen, C. Johnson, and M. Chen. (2013b). Rule-based Visual Mappings—with a Case Study on Poetry Visualization. *Computer Graphics Forum*, 32: 381–390. doi: 10.1111/cgf.12125

Best, S. and S. Marcus. (2009). Surface Reading: An Introduction. *Representations*, 108: 1, pp. 1-21.

Bogan, L. (1968). Night. *The Blue Estuaries: poems 1923-1968*. New York: Farrar, Straus, and Giroux.

Drucker, J. (2013). Performative Materiality and Theoretical Approaches to Interface. *Digital Humanities Quarterly* 7:1 n. pag. www.digitalhumanities.org/dhq/vol/7/1/000143/000143.html (Accessed 26 October 2013).

Montfort, N. Computational Poems. nickm.com/poems (accessed 7 Feb 2014).

Ramsay, S. (2011). *Reading Machines: Toward an Algorithmic Criticism*. Chicago: University of Illinois Press.

Strickland, S. Stephanie Strickland. www.stephaniestrickland.com (accessed 7 Feb 2014).

CatCor: Correspondence of Catherine the Great

Cummings, James
james.cummings@it.ox.ac.uk
University of Oxford

Rubin-Detlev, Kelsey
kelsey.rubin-detlev@mod-langs.ox.ac.uk
University of Oxford

Kahn, Andrew
andrew.kahn@seh.ox.ac.uk
University of Oxford

Introduction

The CatCor pilot project has produced a searchable online text collection of the letters of Catherine the Great. This poster examines the technical background to the project and specifically how the document analysis and subsequent creation of a TEI P5 customization that tightly controlled the more general TEI P5 Guidelines produced a number of significant benefits for the project. These mainly relate both to the ease of checking of the letters and also the subsequent development of a website extracting data from this text collection. This digital humanities project is used as a case study to demonstrate how, with only very minimal funding, a research project may be empowered to produce resources that are of clear and immediate benefit to socio-cultural historians and literary scholars interested in these materials.

About Catherine the Great's letters

Catherine the Great used her correspondence, the primary knowledge-transfer medium of the Age of Enlightenment, to shape her nation's and her own role in the political, cultural, and social arenas of Europe. The collection includes epistolary exchanges with such major figures as Voltaire, Frederick the Great, Friedrich Melchior Grimm, and Catherine's charismatic lover Grigory Potemkin. Catherine ruled Russia for 34 years in the second half of the 18th century (from 1762-1796) and her correspondence was the essential medium through which she governed her court and empire, and established her own and her nation's formidable presence in the political life of Europe and in the intellectual life of the European Enlightenment. Catherine saw herself as the model Enlightenment monarch, ruling over a nation that had fully opened itself to European influence and entered the European consciousness barely 50 years earlier. The primary aim of her reign was to demonstrate, through both policy initiatives and public relations efforts, that Russia was a European state.

The digital collection

Creating a digital collection of this unique set of letters allows scholars to discover unexpected links between letters that have previously been difficult, if not almost impossible, to juxtapose. The use of digital humanities methodologies has meant this project has been able not only to bring together these disparate and difficult to access texts, but also enabled new ways of searching, browsing, and comparing them. While the pilot project only uses 100 letters these were written by Catherine during two key periods in her transformative reign, 1774 and 1790-91. Later funding bids will aim to expand the collection and the project's scope to encompass a complete collection of Catherine's few thousand extant letters and also include those sent to her.

The text of the letters used for the pilot project are encoded in TEI P5 XML and translations to English are provided to facilitate use by a wider range of researchers. This markup also allows the letters to be classified according to a controlled vocabulary of project-specific themes which are then exposed to users as facets in the web interface. The project supplements the letters with a new apparatus of editorial notes and metadata concerning every single person, place, event, and work mentioned in the correspondence. The ease of extraction and aggregation of these named entity instances is one example of the benefits of using a well-known open international standard and having undertaken the necessary document analysis to significantly constrain this schema. Once extracted the instances along with accompanying metadata are able to be displayed, browsed, and filtered with common technologies such as jQuery DataTables. Similarly the use of schematron constraints in the TEI ODD Customization, and the development of straightforward mechanisms for simplifying proofing of common aspects of textual collections is documented in this poster.

Conclusion

The digital humanities aspects of this project are used as a case study into the re-usability of the general purpose tools developed for processing, displaying, and checking basic named entities as well as other common features of such letters. All of the relevant digital technologies, including the TEI P5 XML files of the letters, metadata files, TEI ODD Customization, and XSLT2 functions used for creation and proofing of the underlying data files are available from a public github repository under an open license. It is hoped that this will encourage re-use and empowerment for other digital humanists wishing to make similar text collections available. The open source development work on this project demonstrates the kinds of sophisticated results, immediately beneficial to the scholars working on these materials, that can be achieved with limited resources.

program as animation frames. However, the Kinect sensor has a limitation: if some object (such as another performer, or a body part such as an arm) is placed before a Kinect sensor, detection of the motion is blocked. To overcome limitations of the Kinect sensor in multi-user sensor capture, we used the traditional methods of animation key frames to render the loss movements which cannot be detected. Using thematic material from the original print, we manipulated the motion data to reconstruct the print as a set of animated 3D characters.

When we finished the animation, 3D animated images were added to the original woodcut print. The highlight of the Mazu Festival is the parade with participants depicting historical stories, and new media technology allows viewers to see how performers actually dance and act in the parade, continuing this folk art tradition. Even artist Lin Chih-Hsin, who spent twenty years on his masterpiece, can now see his work in motion. Tradition and folk values, together with an engaging animated display, are integrated to manifest an exciting yet mysterious side of Eastern culture.

Empowering The Matsu(Goddess) Festival Celebration: From Static Woodcut Print to Animated Art

Day, Jia-Ming

jimmyday2010@gmail.com

Taipei National University of the Arts

Hsu, Su-Chu

sucu.hsu@gmail.com

Taipei National University of the Arts

Abstract

The Goddess Mazu, who protects sailors, plays an important role in folk culture in Taiwan and China's southeast coast. In 2009, UNESCO placed Mazu belief and customs on its list of Intangible Cultural Heritage (ICH). Artist Lin Chih-Hsin spent over twenty years making his celebrated woodcut print entitled "Celebrating the Mazu Festival," preserving vanishing rural scenes and simple style of bygone days. The woodcut print is 125 meters long, making it the longest woodcut print by a single solo artist, and it has been presented in a number of exhibitions throughout Europe.

In our work (supported by the [Taiwan] National Science Council), we extended the original static print into animated images. Our goal was to provide a detailed representation of Taiwanese folk art activities and offer opportunities for further research in this field. We used the original style of Lin's woodcut print as a visual theme, recreating characters in 3D in an installation that combined motion sensors, performers, and animated image technology to create a work of animated art. Artist Lin Chin-Hsin was closely involved in this research from the beginning with great enthusiasm, and he has approved the results of the animation.

Constructing the artwork consisted of the following phases: (1) scan an original print from a copy provided by artist Lin; (2) cut out characters from scanned images; (3) build a 3D model of each character; (4) make an empty UV texture map for each model; (5) insert woodcut material texture into the empty UV texture maps; (6) assign UV texture maps back to the 3D model; (7) compare with the original print to ensure accurate representation; and (8) approval by artist Lin of final result.

Our artwork faced two special challenges: (1) animating a static image, and (2) maintaining the original visual style of the print in the animation. Animated content was derived by recording dramatic human performances of folk art groups. Each group acted out a historical story as a dance performed as a parade along the street. From over sixty groups, we chose three on which to perform motion capture using portable Kinect motions sensors to capture individual performer's choreography. The collected motion data was sent to a 3D



Fig. 1: Lin Chih-Hsin "Celebrating the Matsu Festival" exhibition at Staatliches Museum für Völkerkunde München, Germany, 2009.



Fig. 2: The 7th Lord and 8th Lord, who bring the souls of the dead before the judge of the underworld, according to Chinese folk religion.



Fig. 3: Ox fight in spring as an entertainment in rural region.

References

- Lin Chih-Hsin** (2010), *Poseidon Matsu - Lin Chih-Hsin Matsu Festival Celebration Woodcut Prints Watermark Collection*, National Museum of History.
- UNESCO** (2013), *Mazu belief and customs*, Retrieved January 20, from www.unesco.org/culture/ich/index.php?lg=en&pg=00011&RL=00227
- K. P. Aaron Hertzmann** (2000), *Painterly Rendering for Video and Interaction*, Proc. of Non-Photorealistic Animation and Rendering.
- L. M. B. J. M. M. A. K. J. C. L. Robert D. Kalnins** (2002), *WYSIWYG NPR: Drawing Strokes Directly on 3D Models*, SIGGRAPH 2002.
- Jimmy Lin** (2006), *Shape-Oriented Brush Stroke Synthesis in Non-Photorealistic Rendering*, master's dissertation, National Chi Nan University.
- Jun-Lan Yang** (2008), *Simulating the Oil Painting in an NPR 3D Animation System*, master's dissertation, National Chi Nan University.

- Yi-Hsien Chen** (2009), *Synthesizing the NPR Volumetric Visual Effects on Canvas*, master's dissertation, National Chi Nan University.
- A.F.M.C.T.S.M.F.R.M.A.K. a. A.B.J. Shotton** (2011), *Real-time human pose recognition in parts from single depth images*, IEEE CVPR.
- V.G.D.K. a. S.T.C. Plagemann** (2010), *Real-time identification and localization of body parts from depth images*, IEEE ICRA.
- S. S.Park** (2012), *3d hand tracking using Kalman filter in depth space*, EURASIP JASP.
- F.F. a. U.H.M. Baum** (2012), *Tracking ground moving extended objects using RGBD data*, IEEE MFI.
- L.S. a. K.A.M. Luber** (2011), *People tracking in RGB-D data with on-line boosted target models*, IEEE IROS.
- F. Lab.**, *3D Chinese Puppet*, 2005. Retrieved February 21, 2013, from fbi.oddist.org/?page_id=601
- S . C. Hsu** (2005), *Manipulating Digital Archives Using Digital Art Techniques.*, The 4th Digital Archive Conference.
- F. Lab.** (2005), *Explore - Along the River During the Ch'ing-ming Festival*, Retrieved February 21, 2013, from techart.thua.edu.tw/eTaiwan/contents/chingming-index_e.html
- M . Khanna** (2002), *The Crossing*. Retrieved February 21, 2013, from www.sacredworld.info/crossing.htm

Visualizing theatrical heritage: Computer modelling as a tool for researching the theatre history of the Low Countries

De Paepe, Timothy

depaepe.tim@gmail.com

University of Antwerp/Flanders Research Foundation

1. Introduction

In an article, published in 1981, the British theatre scholars C.M. Fogarty and Thomas Lawrenson wrote that the "reconstruction is perhaps the approach in theatre scholarship which comes closest to seizing the essence of the art of theatre."¹ Fogarty and Lawrenson referred to reconstructions of historical performances. However, in my poster presentation I will present a different kind of reconstruction: *virtual* reconstructions of early modern theatre buildings. More specifically, I intend to illustrate how computer models can help theatre and literature historians to better understand and present the vanished theatrical heritage of the Low Countries (present-day Belgium and the Netherlands).

2. Background

Computer modelling and virtual reconstructions are not new. Since John Golder's experimental reconstruction of the seventeenth-century Théâtre du Marais (Paris),² a handful of theatre scholars have, with the help of the computer, reconstructed and visualized historical theatres, with much attention going to early modern (1500-1800) theatres³. Nevertheless, despite the advances in computer modelling and despite 3D modelling software becoming more accessible, such an approach has not been attempted for the early modern theatre history in the Low Countries, even though the approach has a number of potential advantages.

The Low Countries have a rich and dynamic early modern theatre history, including many amateur theatre societies, Chambers of Rhetoric, court theatres, and commercial theatres. Furthermore, the rise and fall of certain theatre evolutions is strongly linked to the tumultuous economic evolutions of the region's cities. Theatre buildings are an externalization of both these theatrical and economic/urban evolutions. Nevertheless,

present-day theatre historiography tells us very little about what these theatre buildings looked like. The economic decline and the lack of world-renowned playwrights have played a role in that, but the lack of (easy to interpret) iconographic evidence is at least as problematic. Often only a combination of information from indirect archival sources, the plays written for these theatres, and ambiguous iconographic material, will tell us more about these places of theatrical performance. Computer models are excellent receptacles for this information, for testing hypotheses and even for studying rejected designs.

The poster presentation will be based on the experiences and results of ongoing research at the University of Antwerp (Literature Department). Part of this research initially focused on creating models of the early modern theatres of Antwerp. Later the research was expanded to include theatres from Ghent (Belgium) and Amsterdam (the Netherlands). So far fourteen theatres or locations of theatrical performances have been reconstructed. I will illustrate my poster with a number of images of these models.

3. Aspects

The poster presentation will include information on the following aspects:

- **Methodology & data:** What sources are available (i.e. what data/input do we have)? What can these sources tell us (i.e. what value does the data have)? What software could be used?
- **Output:** What types of reconstructions can be used, i.e. what can these virtual models look like? I will show both more realistic and visually attractive textured reconstructions on the one hand (fig.1), and more abstract and less hypothetical grey-scale isometric images on the other (fig.2).
- **Advantages:** I wish to address some of the advantages of computer modelling in the field of the theatre history of the Low Countries. For example: virtual reconstructions allow us to show what is lost (both to academic peers and to a general audience); they facilitate any discussion of the theatres; they are "sophisticated tools in primary research—the historians' usual labor of unravelling what the sources do and do not reveal about a performance";⁴ the models help students to dive right into the theatrical heritage, etc.⁵
- **Next steps:** I wish to list a number of potential next steps. These include the creation of a database with standardized, comparable models. But also the reconstruction of less obvious theatrical buildings (e.g. a twentieth-century Dutch synagogue which was later transformed into a community centre/theatre).



Fig. 1: Grand Théâtre, Antwerp, 1711

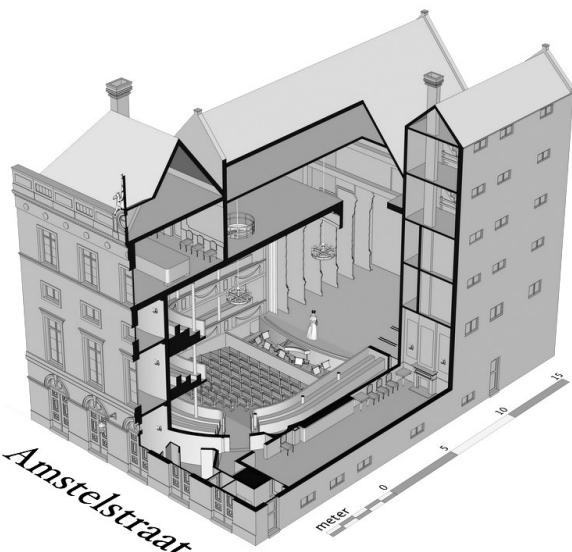


Fig. 2: German Theatre, Amsterdam, ca. 1795

References

1. C.M. Fogarty and Tom Lawrence (1981), *The lessons of the reconstructed performance*, *Theatre Survey* 22, 141-159. Also see: • Robert K. Sariós (1989), *Performance reconstruction: the virtual link between the past and the future*, in Thomas Postlewait & Bruce A. McConachie (eds.), *Interpreting the Theatrical Past. Essays in the Historiography of Performance*. Iowa City: University of Iowa Press), pp. 198-247.
2. John Golder. (1984) *The Théâtre du Marais in 1644: a new look at the old evidence concerning France's second public theatre*, *Theatre Survey* 25, pp.127-152.
3. David Thomas (1999), *The design of the Théâtre du Marais and Wren's Theatre Royal, Drury Lane: A computer-based investigation*, *Theatre Notebook* 53:3, pp. 127-145; Christa Williford (2001), *Computer Modelling Classical French Theatre Spaces: Three Reconstructions*, in Philip Tomlinson (ed.), French 'Classical' Theatre Today. Teaching, Research, Performance (Amsterdam: Rodopi), pp. 155-164; Frank Mohler (1999a), *The Survival of the Mechanized Flat Wing Change: The Court Theatres of Gripsholm, Cesky Krumlov, and Drottningholm*, in *Theatre Design & Technology* (1999a): 6-56; Frank Mohler (2004), *The Chateau Theatre in Litomysl and the Scenery of Josef Platzer*, in *Theatre Design & Technology* (2004): 24-31.
4. David Thomas (1999), *The design of the Théâtre du Marais and Wren's Theatre Royal, Drury Lane: A computer-based investigation*, *Theatre Notebook* 53:3, pp. 127-145; Christa Williford (2001), *Computer Modelling Classical French Theatre Spaces: Three Reconstructions*, in Philip Tomlinson (ed.), French 'Classical' Theatre Today. Teaching, Research, Performance (Amsterdam: Rodopi), pp. 155-164; Frank Mohler (1999a), *The Survival of the Mechanized Flat Wing Change: The Court Theatres of Gripsholm, Cesky Krumlov, and Drottninghol*, in *Theatre Design & Technology*: 6-56; Frank Mohler (2004), *The Chateau Theatre in Litomysl and the Scenery of Josef Platzer*, in *Theatre Design & Technology*: 24-31.
5. Richard Beacham (1999), 'Eke out our performance with your mind': reconstructing the theatrical past with the aid of computer simulation, in Terry Coppock (ed.), *Information Technology and Scholarship. Applications in the Humanities* (London: The British Library), pp. 131-154; Frank Mohler (1999), *Computer Modelling as a Tool for the Reconstruction of Historic Theatrical Production Techniques*, *Theatre Journal* 51:4, pp. 417-431.

Orthography and Biblical Criticism

Dershowitz, Idan

Tel Aviv University, IL

Dershowitz, Nachum

Tel Aviv University, IL

Hasid, Tomer

Tel Aviv University, IL

Ta-Shma, Amnon

Tel Aviv University, IL

Abstract

Biblical Hebrew exhibits considerable orthographic variability. A single word may be spelled in multiple ways—often within the same book. In this study, we set out to determine if these differences in spelling correspond in any way to scholarly theories regarding the authorship of the Pentateuch. We use a statistical test that is designed for use when there are many features to take into account, each of which occurs only sparsely. Our results indicate that despite the tortuous editing processes and countless generations of hand-copied manuscripts, certain statistically significant correlations between orthography and the hypothesized sources remain.

1 Introduction

The Pentateuch has been attributed to several major sources. We investigate here whether there exists a statistically significant correlation between these postulated sources and variations in spelling in the received Masoretic text.

We consider the source units about which there is broad agreement among Bible scholars, namely, the classic four-source division of the text into J, E, P, and D¹, plus the consensual source H. We only considered words occurring in paragraphs for which there is relative agreement among scholars. We also compared genres—narrative and legal—since different genres might employ different conventions. (We ignored poetry with its distinctive language register.)

Regarding orthography, we examined the use of consonants to represent vowels, a practice that has changed over time. The Canaanite languages—Hebrew among them—were generally recorded in alphabetic writing sans vowels. With time, certain characters began to serve double duty, representing vowels, as well as consonants. These letters are known as matres lectionis ("mothers of reading"). The written representation of vowels increased from one century to the next, but it appears there was variation even within a single period.

While matres lectionis proliferated, another process complicated matters. When a word's pronunciation evolved so that a particular consonant stopped being pronounced, the letter representing that consonant was not always written. For this reason, we classify spellings as either "neological" (reflecting innovative orthography) or "paleological" (conforming to earlier norms).

We have, then, two labelings to work with:

- By source and genre: J, E, E-law, P, P-law, D, D-law, H, or H-law. For simplicity, we will refer to these nine categories as "sources."
- By orthography: paleological or neological.

We apply a statistical test, Cochran-Mantel-Haenszel (CMH)², to check whether there is a correlation between the two labelings, that is, whether any particular source is more paleological than others.

2 Possible Approaches to the Statistical Problem

Assume we have two sources, A and B, plus an orthographic classification, and would like to check whether the classifications are correlated.

2.1 The Naïve Approach

A naïve approach is to count the total frequencies of neological and paleological syllables for each one of the sources and then run a χ^2 test for the resulting 2×2 table. We believe this approach is not good. If A has a different word distribution than B, it is possible that even when the sources have identical spelling the naïve test would declare the two classifications strongly correlated, simply because one tends to use certain words that are spelled as neological more often. Working with aggregated data is therefore most likely to catch word distribution differences between sources, rather than spelling differences.

2.2 Filtering

Andersen and Forbes³ conducted an extensive automated study of spelling in the Bible. They created 65 classes, based on grammatical form, vocalization and stress. Within each, they used the naïve approach, aggregating all words in a class and checking the χ^2 score of the A/B and plene/defective classifications. As they use aggregated data (within each class), they still face the word distribution problem, and in particular their method is vulnerable to words like ḥōdō (lo) that appear frequently and mostly in one form. They tackled this problem with several ad-hoc filters and rules, such as filtering out words that almost always appear only in one form (see, e.g., [1, Chap. 10]). However, on a conceptual level, it seems that whatever set of filters is used, there is still the problem that differences in word distribution between different sources is interpreted as spelling differences. It appears they could not exhibit a conclusive relationship between stress and spelling (see [1, Epilog]), which seems to undermine the rationale behind dividing words into those classes.

2.3 New Approach

Our goal is to identify spelling differences even when each source may have a different distribution of words (e.g., legal texts tend to use legal terminology). The appropriate statistical technique is CMH. The idea is to bypass the language distribution problem by having a 2×2 contingency table for each word in the language, describing the number of neological/paleological occurrences of the word in each source. The test combines the data from all the 2×2 tables in a way that gives weight to the statistical significance of the data in each table, ignoring the frequencies of the word in each source.

We enumerate events at the finest possible granularity, classifying each syllable of each occurrence of a particular word in the text. For each syllable, we have one stratum (in the statistical sense of stratified data) containing a 2×2 contingency table describing the number of neological/paleological occurrences of that syllable of the word in each source.

This observation, as simple as it is, is conceptually important and is crucial for getting sound statistical data on the problem. As a side effect of using CMH, we avoid ad-hoc filters and rules.

3 Experimental Design and Results

We have a stratum for each syllable and use the tagging of the biblical text into word senses (Strong number). We consider two sources at a time, computing the following statistics:

1. the χ^2 and p-value of the CMH test;
2. the validity of the χ^2 test with the Rule of 5;
3. the common odds ratio;
4. the $p=1-\alpha$ confidence intervals for the logarithm of the common odds ratio, taking $\alpha = 0.05$.

The p-values and the ln(odds) values for the pairs of sources are tabulated as follows:

	D	D-law	E	E-law	P	P-law	H	H-law
D-law	0.900							
E	0.073	0.000						
E-law	~	0.198	~					
P	0.323	0.848	0.777	~				
P-law	0.000	0.087	0.588	~	0.327			
H	~	~	~	~	0.445	~		
H-law	0.296	0.804	0.482	~	0.240	0.067	~	
J	0.108	0.033	0.671	0.790	0.852	0.276	~	0.184

The cells with tildes are those that failed to pass the Rule of 5 [3].

In the following table, the number in cell (i, j) tells us how much source i is more likely to be paleological than source j . Roughly speaking, if the cell $(i, j) = 0.44$, then i uses the paleological form 20.44 ≥ 1 more often than j . If it is zero they have the same frequencies, 20 = 1; if it is negative, source i is less paleological, 2–0.44 ≤ 1 .

	D	D-law	E	E-law	P	P-law	H	H-law
D-law	-0.080							
E	0.460	1.192						
E-law	~	0.776	~					
P	0.263	0.118	-0.087	~				
P-law	0.818	0.503	0.210	~	0.181			
H	~	~	~	~	0.546	~		
H-law	0.516	0.161	-0.388	~	-0.389	-0.538	~	
J	0.351	0.521	0.107	-0.237	0.054	0.267	~	0.653

Thus, D-law appears to be the most neological.

4 Discussion

Our results appear to be of potential interest for several reasons.

- For Bible scholars, they suggest that the countless scribes who edited, expanded, and copied the text(s) that eventually crystallized into the Masoretic text did not change enough to obscure the characteristic spelling of individual units. Our findings open the door to new approaches in the critical analysis of biblical texts, as the value of orthography in such contexts has thus far been underestimated.
- The simple statistical test we use cannot disentangle the many authors of the Bible. However, it does produce some interesting results, that we hope would be combined with other data to shed light on the fascinating question of how the Bible, as we know it today, evolved. In particular, the observation that Deuteronomic narrative is more neological in spelling than Priestly law may be of some value in the ongoing debate regarding the relative dating of P and D. It is possible that there is some hidden random variable that strongly affects spelling, which might better explain the results. However, in spite of much effort (e.g., see [1]), such a hidden random variable has not been identified.
- Outliers in the analysis may suggest alternate classifications for some linguistic phenomena. The most prominent example for that is the holam syllable פַּה (po) in the word צִפּוֹר (tzipor). The experts labeled צִפּוֹר as paleological and פַּה as neological, while our data seems to indicate the opposite (in D-law there are 2 occurrences of צִפּוֹר and none of פַּה, in P-law 2 occurrences of פַּה and 9 of צִפּוֹר, though D-law is overall more neological than P-law).
- For textual analysis, more generally, this work suggests that the Cochran-Mantel-Haenszel is an appropriate statistical measure, when features are sparse.

References

1. S. R. Driver (1913). *An Introduction to the Literature of the Old Testament*. Edinburgh.
2. N. Mantel and J. L. Fleiss (1980). *Minimum expected cell size requirements for the Mantel-Haenszel one-degree-of-freedom chi-square test and a related rapid procedure*. American Journal of Epidemiology, 112(1):129–134.

3. F. I. Andersen and A. D. Forbes (1986). *Spelling in the Hebrew Bible*. Biblical Institute Press.

L'Innommable / The Unnamable: The Second Module of the Samuel Beckett Digital Manuscript Project's Hybrid Genetic Edition.

Dillen, Wout

wout.dillen@gmail.com

University of Antwerp

This poster will offer an interactive demonstration of the second module of the *Samuel Beckett Digital Manuscript Project* (BDMP) – an international collaboration between the Centre for Manuscript Genetics at the University of Antwerp, the Beckett International Foundation at the University of Reading, and the Harry Ransom Humanities Research Center at the University of Austin, Texas, with the kind permission of the Estate of Samuel Beckett. As its name implies, the BDMP aims to reunite and make publicly accessible all manuscripts of Samuel Beckett's works, the physical documents of which are located in different holding libraries around the world. This goal will be realized in the form of a hybrid genetic edition that combines a digital archive of the manuscripts organized in twenty-six research modules (one published each year) with a series of twenty-six accompanying volumes analysing the geneses of the texts contained in the corresponding modules. Each of the modules comprises digital facsimiles and transcriptions of all extant manuscripts pertaining to an individual text – or to a collection of shorter texts. The digital archive can be accessed online at www.beckettarchive.org, where you can currently find the project's first module, which combines the manuscripts of the late short prose text *Stirrings Still / Soubresauts* with those of Beckett's last poem *Comment dire / what is the word*¹. The project's second module, which will present the bilingual genetic dossier of Beckett's novel *L'Innommable / The Unnamable* edited by Dirk Van Hulle, Shane Weller and Vincent Neyt, will be made available online towards the end of 2013².

As a hybrid genetic edition, the BDMP combines digital scholarly editing with genetic criticism – a form of literary criticism that studies the dynamics of the writing process. As such, the edition does not aim to support a new reading text of Beckett's works, but rather to highlight the creative process that brought those works about. Traces of this process can be found in the extant manuscripts, more specifically in their many deletions, additions, paralipomena, doodles, etc. According to Pierre-Marc de Biasi, the objective of genetic criticism is twofold (42)³: first (1) to locate, collate, and transcribe all of the work's extant versions in order to make them analysable, and then (2) to reconstruct the logic of the work's genesis (also called the 'avant texte') from a chosen perspective. Therefore, rather than focussing solely on an *analysis* of the works' geneses – which can be found in the interpretative, printed component of our edition – we also want to make these geneses *analysable*, by offering fully transcribed (in TEI-compliant XML) and searchable facsimiles in the edition's digital component. Hence, the BDMP complies with Patrick Sahle's recent definition of scholarly hybrid editions, stipulating that such editions should not only be published in different media, but that their different components should complement one another, and that each component should take the possibilities and limitations of its medium into account (64)⁴.

Because the modern manuscripts we exhibit are still under copyright and therefore do not yet belong to the public domain, the BDMP still requires its users (either individuals or institutions) to pay a subscription fee to gain access to its materials. Therefore, this poster will be a great opportunity for potential users to receive a personal, hands-on introduction to the project. The poster's main focus will be on the project's

newest module, and on the differences between both modules. Because the genesis of every work is different, even within the oeuvre of a single author, we try to determine what the specific needs of the texts in each subsequent module are, and to re-evaluate the tools and functionalities the module will provide to satisfy those needs accordingly. For the BDMP's second module, for example, this resulted in changes to the Synoptic Sentence View (which allows the user to grab any sentence in our corpus, and generate a chronological list of its different versions), to the image-text linking tool, etc. Furthermore, this poster will also demonstrate the BDMP's improved integration of CollateX⁵ – the interoperable collation tool that is part of the Interedition project⁶.

References

1. Beckett, Samuel (2011). *Stirrings Still / Soubresauts and Comment Dire / what is the word: an electronic genetic edition* (Series 'The Beckett Digital Manuscript Project' module 1), edited by Dirk Van Hulle and Vincent Neyt. Brussels: University Press Antwerp (ASP/UPA). www.beckettarchive.org (accessed 30 October 2013).
2. Beckett, Samuel (2011). *Stirrings Still / Soubresauts and Comment Dire / what is the word: an electronic genetic edition* (Series 'The Beckett Digital Manuscript Project' module 1), edited by Dirk Van Hulle and Vincent Neyt. Brussels: University Press Antwerp (ASP/UPA). www.beckettarchive.org (accessed 30 October 2013).
3. de Biasi, Pierre-Marc (2004) *Toward a Science of Literature: Manuscript Analysis and the Genesis of the Work*. Genetic Criticism. Texts and Avant-textes. Philadelphia: University of Pennsylvania Press: 36-68.
4. Sahle, Patrick (2013) *Digitale Editionsformen, Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandelns*. Teil 2: Befunde, Theorie und Methodik. Nordstedt: Books on Demand.
5. CollateX (2013) *Home*. CollateX. www.collatex.net (accessed 30 October 2013).
6. Interedition (2013) *Home*. Interedition. www.interedition.eu (accessed 30 October 2013).

Spreading DiRT: extending the Digital Research Tools directory

Dombrowski, Quinn

quinn@berkeley.edu
UC Berkeley

Gold, Matthew

mattgold@gmail.com
City University of New York

1. Background

The DiRT (Digital Research Tools, dirt.projectbamboo.org) directory is a longstanding resource for scholars interested in digital tools and methodologies, providing basic information about software that can facilitate different stages of the research process. DiRT was originally designed as a wiki, where a single wiki page contained information about all tools in a given category. In 2011, under the auspices of Project Bamboo, DiRT was completely rebuilt using the Drupal content management system, which allowed for data to be stored in a structured manner. This enabled more complex searching and browsing options (such as allowing the user to limit results based on criteria like platform or cost), and provided individual profile pages for each tool, which could then serve as a locus for specific comments, or be referenced in other tool profiles. For instance, if a profile page indicates that Neatline is a suite of add-on tools for Omeka, a link to Omeka appears on the Neatline tool profile page, and vice versa.

2. Current development project

One of the biggest limitations of DiRT has been the fact that its contents-- the product of a considerable amount of volunteer work-- have only been available via DiRT's own web interface. Creating and curating the tool listings on DiRT is largely a manual process. A steering and curatorial board takes an active role in shaping the ongoing development of the site and ensuring data quality, but individual contributions by users make up a large portion of the data. DiRT is currently undergoing a new phase of development, supported by the Andrew W. Mellon Foundation, with the goals of making DiRT data available to others who want to incorporate information about tools into other projects, resources and environments, and also expanding the content provided by DiRT to more clearly situate the tools in the contexts of the projects, research workflows, and pedagogical activities that use them. This poster will demonstrate the accomplishments of the current development project and include information about opportunities to get involved with the project, by trying the DiRT API and plug-ins, or contributing to tool reviews and documented workflows.

3. Areas of work

The poster will also highlight the progress made on developing a DiRT plug-in for Commons In A Box (CBOX), an open source scholarly networking platform created by the City University of New York and used by the Modern Language Association (MLA), the regional NYC Digital Humanities group, and an increasing number of projects and organizations that could benefit from integrated access to information about tools. The CBOX plug-in will:

1. provide users with the ability to display information about their DiRT site activity (e.g. tool contributions and edits, reviews, and tool usage information) on their Commons profile;
2. provide an interface for searching DiRT within the CUNY Academic Commons, for use by groups with an interest in digital humanities;
3. provide a link to DiRT to facilitate access for inputting new tools.

The poster will also illustrate other areas of development including:

- Use of the DiRT API and the API for the DHCommons project directory to augment DiRT tool profiles with information about what projects are using a particular tool
- Guidelines and examples of best practice for writing tool reviews, with potential pedagogical applications (e.g. providing a framework for instructors who want to assign students to write reviews of digital research tools, which could then be refined for ultimate publication on DiRT)
- Guidelines and examples of best practice for documenting workflows or "recipes" that combine multiple tools in the DiRT registry to achieve some research objective.
- Adoption of the taxonomy of research methods jointly developed between DiRT and DARIAH-DE, to replace the previous ad-hoc set of tool categories. DiRT will serve as one of three initial test cases for this taxonomy, which has benefitted from extensive public feedback.
- Documentation for how to develop custom tool lists (e.g. tools to be used in a particular class, or tools that are particularly relevant for the disciplines that a subject specialist librarian supports) that pull from the information stored in DiRT, and display that information on other sites.

References

- Babeu, Alison** (2006). *Rome Wasn't Digitized in a Day": Building a Cyberinfrastructure for Digital Classicists*. CLIR reports, August 2011.
Borgman, Christine L. "The Digital Future Is Now: A Call to Action for the Humanities." *Digital Humanities Quarterly* 3, no. 4 (Fall 2009). <http://www.digitalhumanities.org/dhq/vol/3/4/000077/000077.html>.
Unsworth, John. Our Cultural Commonwealth: The Report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences. American Council of Learned Societies. http://aclss.org/uploadedFiles/Publications/Programs/Our_Cultural_Commonwealth.pdf.

Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences. American Council of Learned Societies. http://aclss.org/uploadedFiles/Publications/Programs/Our_Cultural_Commonwealth.pdf.

An easy tool for creating digital scholarly editions

Dumont, Stefan

dumont@bbaw.de
 Berlin-Brandenburg Academy of Sciences and Humanities

Fechner, Martin

fechner@bbaw.de
 Berlin-Brandenburg Academy of Sciences and Humanities

The Berlin Brandenburg Academy of Sciences and Humanities (BBAW) is home to multiple long term research projects which encompass various fields of study. The research group TELOTA (The Electronic Life of the Academy) supports the digital humanities aspects of these projects, including developing software solutions for the daily work of their researchers.

Experience shows that although using TEI-XML encoding¹ for the digital transcription and annotation of manuscripts can improve research in edition projects, the readiness to implement it greatly relies on the user-friendliness of the entry interface. From the perspective of a researcher, working directly in XML simply doesn't compare to the ease of programs like MS Word. A new software solution must therefore at least offer the same amount of editorial comfort as such programs. Ideally, it would also encompass the complete life-cycle of an edition: from the first phases of transcription to the final publication.

TELOTA developed such a software solution "ediarum"², which can be adapted to the needs of different research projects. The solution consists of various software components that allow the researchers to construct and edit transcriptions of manuscripts and manuscript descriptions into XML following the TEI guidelines. Through editing and tagging manuscripts, a large quantity of digital data can be generated and is prepared for further research and presentation. The solution includes the possibility to create apparatuses of different kinds, as well as to create without much additional effort both a print and web publication.

The central software component of the new digital work environment "ediarum" is Oxygen XML Author³. The researcher does not edit the XML code directly, but instead works in a user-friendly Author mode, which is designed through Cascading Stylesheets (CSS). The researcher is able to choose more than one perspective within the Author view, and thus can select per mouse click the appropriate perspective for the current task. Additionally, a toolbar is provided with which the researcher can enter markup with the push of a button. In this way text phenomena such as deletions or additions, or editorial commentary, are easily inserted. Person and place names can also be recorded with their appropriate TEI markup, and in addition they can be simultaneously linked to the corresponding index. This is done through selecting the name from a convenient drop down list. The entire manuscript text can thus be quickly and simply marked up with TEI conform XML. All documents are stored in the open source database "existdb".⁴

Besides creating a digital work environment for "ediarum" in Oxygen XML Author, a website can also be built based on eXistdb, XQuery, and XSLT. Through the website the researchers can easily page through or search the current data inventory.

Through the intergration of ConTeXt⁵ into "ediarum" a further publication type, a print edition, can automatically be generated as a PDF from the TEI XML document. The layout and format imitates previously printed volumes of critical editions. Each TEI element must be given a specific formatting command through a configuration file. In this way the different apparatuses appear

as footnotes that refer to the main text with the help of line numbers and lemmata. The print edition can also provide the suitable index for each transcription and solves any occurring cross references between manuscripts.

This work environment has been in use since 2012 by the staff of different research projects for their daily work. When asked their opinion, the researchers were in agreement that the new work environment greatly eased their editorial work and saved them significant time. The possibility to directly check the results of their work in a web presentation or as a printed edition was seen as an additional positive aspect.

Presently "ediarum" is used by the edition of letters, lectures and diaries of the German theologian Friedrich Schleiermacher⁶; the description of medieval manuscripts of Aristoteles or the regesta of emperor Friedrich III. For each project the TEI XML schemata and main functions were customized to the different manuscript types and needs. In the future, "ediarum" will be implemented for further projects of the academy or other institutions, e.g. the edition of Jeremias Gotthelfs prints and manuscripts at the University of Bern (Suisse).

Screenshots

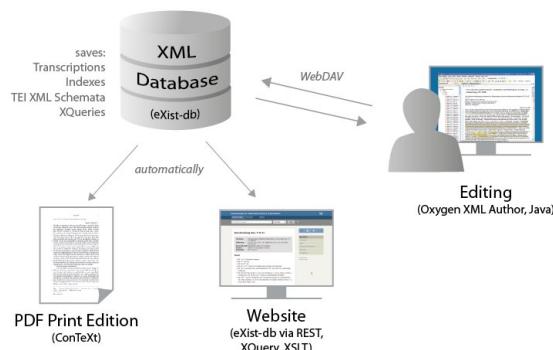


Fig. 1: Workflow of ediarum

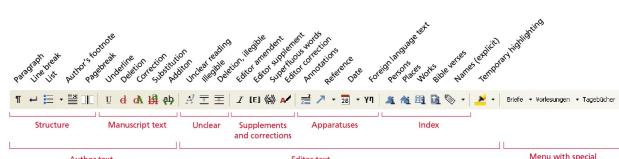


Fig. 2: Customized TEI toolbar in Oxygen XML Author for the digital edition of Friedrich Schleiermachers letters, lectures and diaries

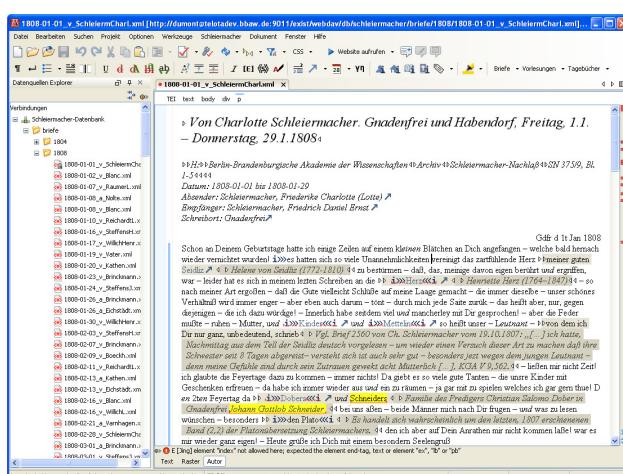


Fig. 3: Transcription of a letter in Oxygen XML Author

References

1. Burnard, Lou; Bauman, Syd (Hg.): *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Charlottesville, Virginia, USA 2007. URL: www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf
2. www.bbaw.de/en/telota/software/ediarum
3. User Manual Oxygen XML Author 14. URL: www.oxygenxml.com/doc/ug-editor/
4. eXist Main Documentation. URL: www.exist-db.org/exist/documentation.xml
5. ConTeXt Dokumentation. URL: wiki.contextgarden.net/Main_Page
6. Dumont, Stefan; Fechner, Martin: *Digitale Arbeitsumgebung für das Editionsvorhaben »Schleiermacher in Berlin 1808–1834* In: di giversity — Webmagazin für Informationstechnologie in den Geisteswissenschaften. URL: digiversity.net/2012/digitale-arbeitsumgebung-für-das-editionsvorhaben-schleiermacher-in-berlin-1808-1834

Créer un centre de recherche interuniversitaire sur les humanités numériques au Québec : Défis et succès

Eberle-Sinatra, Michael

Université de Montréal, Canada

Sinclair, Stéfan

McGill University, Canada

Dyens, Olliver

McGill University, Canada

Vitali Rosati, Marcello

Université de Montréal, Canada

Notre poster portera sur la création du Centre de recherche interuniversitaire sur les humanités numériques (CRIHN), les étapes ayant menées à sa création, surtout liées aux difficultés inter-institutionnelles et en relation avec les organismes subventionnaires, plus la question de la langue française dans un contexte des humanités numériques à l'échelle mondiale. Notre poster sera donc une ressource d'information pour d'autres chercheurs qui pensent mettre en place un nouveau centre en présentant notre expérience et en la mettant en perspective avec d'autres centres canadiens et internationaux.

Le CRIHN est basé à l'université de Montréal dans des locaux multimédia qui permettront des rencontres mensuelles en personne et en vidéo- conférence simultanée pour les membres hors Montréal. La direction du centre est assumée par un professeur de l'université de Montréal (Michael E. Sinatra) pour les quatre premières années de son existence, avec l'assistance d'un directeur adjoint a McGill (Stéfan Sinclair) et d'un autre a Concordia (Jason Camlot). La direction ira ensuite en rotation pour un période de quatre ans dans ces deux autres institutions.

Le poster contiendra aussi des informations sur la constitution du CRIHN et la manière dont nous avons structuré les axes de recherches avec nos chercheurs et les institutions impliquées. Le CRIHN qui regroupe 31 chercheurs québécois à la pointe de la recherche sur la culture numérique. Provenant d'horizons disciplinaires variés mais partageant une expérience théorique et pratique, le centre a voulu mettre en avant l'expertise en humanité numérique répartie au Québec à travers sept universités, CEGEP et établissement de recherche et d'enrichir les collaborations présentes et futures en créant une synergie interdisciplinaire et interuniversitaire.

Sous la direction Michael E. Sinatra, la recherche menée dans le CRIHN s'articule autour de deux grands axes. Le premier, intitulé 'Théorie(s) du numérique' (sous la direction d'Olliver Dyens) regroupe des chercheurs qui développent une réflexion théorique sur les enjeux du numérique et sur les changements culturels déterminés par les technologies. L'objectif de cet axe est de fédérer les travaux d'analyse de

ces chercheurs afin de clarifier certaines notions clés qui caractérisent l'environnement numérique. Le travail collectif des chercheurs vise à produire des définitions terminologiques et conceptuelles sur lesquelles pouvoir ensuite construire un langage commun et des principes théoriques partagés. Ce travail sert aussi à rendre plus clairs les objectifs plus pratiques des autres volets de recherche du centre.

Le deuxième axe, intitulé 'Les outils du savoir' (sous la direction de Stéfan Sinclair) réunit plusieurs chercheurs très actifs dans le développement et l'utilisation d'outils informatiques conçus pour la création, l'analyse, la visualisation et la diffusion du savoir dans les humanités. L'ambition principale des chercheurs de cet axe est de donner un coup de pouce décisif aux transformations médiatiques en cours dans les recherches des sciences humaines, du texte imprimé toujours dominant vers le support numérique et tout son potentiel interactif et analytique.

Les activités du CRIHN sont organisées autour des projets actuels et futurs des membres du centre et prendront diverses formes d'activités de diffusion (colloques et ateliers, table-rondes avec les média, expérimentations avec de nouvelles formes de publications électroniques, etc.) ainsi que des activités de formation des étudiants, une composante importante des activités du centre.

Le CRIHN offre aussi, dans un axe transversal, un observatoire (sous la direction de Marcello Vitali Rosati) qui développe une veille sur les pratiques émergentes et une vulgarisation des approches liées à la culture numérique. Son objectif principal est donc celui de communiquer avec le grand public pour que la société puisse avancer dans sa façon d'interpréter le rôle et les enjeux de la culture numérique. Le CRIHN permet donc à travers son observatoire de mettre en avant une expertise québécoise reconnue nationalement et internationalement dans le monde académique mais sans vitrine appropriée pour le milieu journalistique et les citoyens.

Stylometry, network analysis, and Latin literature

Eder, Maciej

maciejeder@gmail.com

Pedagogical University of Krakow

St Jerome, early-Christian writer and translator of the Bible, claims that he had a dream in which the God accused him: "You are a Ciceronian, not a Christian!", because he had paid too much attention to the beauty of the Ciceronian style. St Jerome's dream reflects Christian antiquity's general attitude to classical literature: pagan texts were claimed to be generally dangerous, and thus they were rarely imitated. Centuries later, Renaissance humanists "discovered" the classical authors again, and they intended to purge Latin of medieval traces. There was no clear answer how to restitute the Latin style, though. The discussion about the ways of imitation, known as the Ciceronian Quarrel, was the single most important literary debate of the Renaissance (DellaNeva, 2007)¹. Arguably, all these changes in the attitude to classical literature were followed by style changes measurable with stylometric methods.

While computational stylistics has been usually associated with authorship attribution, recent research shows that the same methods can be used in a much broader context of literary study. Namely, the underlaying idea of tracing similarities between (anonymous) texts can be extended to map textual relations in large-scale approaches to literature. Additionally, they are supported by the seminal concepts of “computation into criticism” (Burrows, 1987)², “distant reading” (Moretti, 2013)³ and “macroanalysis” (Jockers, 2013)⁴.

In the present study, stylometric techniques are combined with network analysis to reveal the strongest similarities between analyzed texts on the one hand, and some deeper or less obvious relations, usually filtered out by standard

nearest neighbor methods, on the other (Eder, forthcoming)⁵. Consensus network visualizations (cf. Fig. 1) attempt to overcome drawbacks of Cluster Analysis and other explanatory multidimensional techniques. Firstly, these methods are highly dependent on the number of features (e.g. word frequencies) analyzed, secondly, they are not suitable to visualize large datasets. The technique applied in this study counts frequencies of frequent words in a corpus, computes distances (differences) between samples, and for each sample identifies its nearest neighbors (i.e. the samples that turned out to be the most similar). Next, neighboring samples are turned into connected nodes of a network. The above procedure is repeated many times, in each iteration a different range of frequent words is analyzed. Finally, particular networks produced for each iteration are combined into a single consensus network.

A few dozens of ancient, medieval, and early modern Latin texts (prose and poetry) were harvested from different open-access databases, and analyzed using consensus networks. The research problems to be undertaken included: (1) an analysis of style variation in the classical Latin prose of the Augustan Age and the Silver Age (Cicero, Caesar, Tacitus, Livy, Suetonius, Apuleius, etc.). The aim was to identify which features of language are author-dependent, and which are affected by the literary epoch; (2) an investigation of the Renaissance “restitution” of classical Latin (in a Ciceronian flavor), as opposed to medieval Vulgar Latin as introduced by early Christian writers (Bolgar, 1954)⁶. This was focused on the question of the extent to which the Renaissance humanists succeeded in imitating the style of Cicero, and whether they truly overcame the medieval vulgar style (as they claimed to have done). Last but not least, (3) an examination of the problem of “Attic” prose and the anti-Ciceronian movement of the late Renaissance and early Baroque, based on the analysis of the style of Justus Lipsius, Erasmus and other writers: Puteanus, Moretus, Fredro etc. (Croll, 1924, 1996; Salmon, 1980; Tunberg, 1999)^{7 8 9 10}. The questions were as follows: is the trace of Seneca’s and Tacitus’ style indeed noticeable in this modern “Attic” way of writing? Did the “Attic” authors really escape from Ciceronianism in style?

One of the stylometric consensus networks is shown in Fig. 1. It visualizes the textual similarities between 55 poetic texts, in order to assess the question of sequels in Latin poetry. Highlighted texts include Maffeo Vegio's continuation of the *Aeneid*, Thomas May's *Supplementum Pharsaliae*, and John of Garland's *Integumenta super Ovidium Metamorphoseos*, in comparison with their ancient counterparts written by Virgil, Lucan, and Ovid, respectively. One can clearly notice that both Vegius and May were quite successful in imitating the classical authors, while Garlandus's traces to Ovid are significantly weaker.

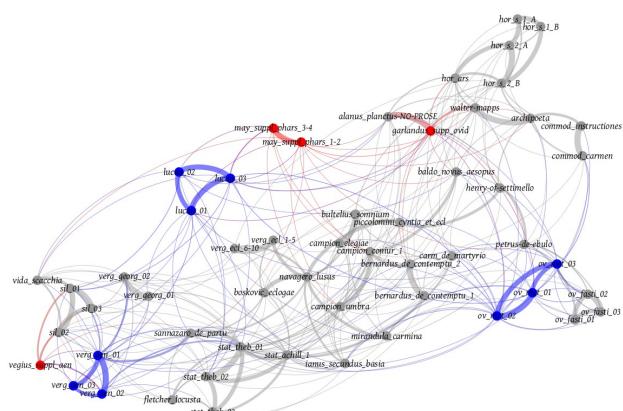


Fig. 1: Sequels in Latin poetry: network of 55 poetic texts. Imitated poems marked in blue, their followers marked in red.

Other networks were produced for prose. Standard network analysis procedures were applied, and the data were visualized using consensus networks. Instead of expected

chronological patterns, the results revealed some other interesting regularities.

Analyzing Latin style with stylometric methods, one should remember that the medieval authors relatively often cite the Bible and related sources, while the humanists' treatises are full of explicit and/or implicit quotations from classical literature. More importantly, the humanists consciously tried to avoid medieval vocabulary in favor of words that were used at least once by Cicero. For that reason, a stylometric comparison of medieval and early modern Latin brings some additional issues, intensive text re-use being one of the most important (Eder, 2013). Some of these issues will be discussed in detail.

References

1. DellaNeva, J. (2007). *Ciceronian Controversies*. Cambridge, MA: Harvard University Press.
2. Burrows, J. (1987). *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method*. Oxford: Clarendon Press.
3. Moretti, F. (2013). *Distant Reading*. London: Verso.
4. Jockers, M. (2013). *Macroanalysis: Digital Methods and Literary History*. Champaign: University of Illinois Press.
5. Eder, M. (2014). *Visualization in stylometry: some reliability issues* (forthcoming).
6. Bolgar R. R. (1954). *The Classical Heritage and its Beneficiaries*. Cambridge: Cambridge University Press.
7. Croll M. W. (1924). *Muret and the History of "Attic" Prose*. PMLA, 39: 254–309.
8. Croll M. W. (1996). "Attic" prose in the seventeenth century. In: Style, Rhetoric and Rhythm. Ed. by J. M. Patrick and R. O. Evans. Princeton NJ: Princeton University Press, pp. 51–101.
9. Salmon, J. H. M. (1980). *Cicero and Tacitus in Sixteenth-Century France*. The American Historical Review, 85: 307–31.
10. Tunberg, T. O. (1999). *Observations on the style and language of Lipsius's prose: a look at some selected texts*. In: Tournoy, G., Landsheer, J. de, and Papy, J. (eds), *Iustus Lipsius: Europae lumen et columen*. Leuven: Cornel University Press, pp. 169–78.
11. Eder, M. (2013). *Mind your corpus: systematic errors in authorship attribution*. Literary and Linguistic Computing, 28(4): 603–14.

Network Analysis for the History of Religions – The SeNeReKo project

Elwert, Frederik

frederik.elwert@rub.de
Ruhr-University Bochum

Wortmann, Sven

sven.wortmann@rub.de
Ruhr-University Bochum

Hofmann, Beate

beate.hofmann@rub.de
Ruhr-University Bochum

Knauth, Jürgen

knauth@uni-trier.de
Trier University

The SeNeReKo project is a joint research project of the Center for Religious Studies (Bochum, Germany) and the Center for Digital Humanities (Trier, Germany), funded by the German Federal Ministry of Education and Research. The project develops methods for the application of network analysis as a tool for the study of historical religious texts. Informed by theoretical approaches in the study of religion which stress the importance of interreligious contacts for the formation and development of religious traditions, concepts from network theory are used in order to analyse the context and interdependence of religious actors and concepts.

Religious traditions are seen as developing face to face of each other, developing their own meaning systems by at the same time recognising and rejecting – or adopting and transforming – meaning systems of "others".¹ The relational framework of network analysis is taken as a methodological tool to operationalise such approaches.

The analytical process consists of two steps: Firstly, networks have to be created on the basis of the textual sources by identifying nodes and their relations. In contrast to applications of network analysis that are only interested in the relations of (historical or literary) social actors,² the project also takes semantic structures into account. During the first step, methods from computational linguistics (e.g. part-of-speech tagging, lemmatisation, syntax parsing, anaphora resolution) are applied in order to identify relevant terms (nodes) and their relations (edges). Since such tools are not available for the target languages, existing state-of-the-art tools are adapted for these tasks. Building on these linguistic annotations, networks of words (terms/lemmas) are created. For this purpose, a set of tools is built, implementing several existing and newly developed network creation algorithms.³ These tools are compatible to the CLARIN WebLicht infrastructure, using TCF as their data exchange format, and ISOcat for handling linguistic tag sets. They will be made available under an open source license.

Secondly, the resulting networks are analysed using methods from network analysis. During this step, centrality measures and clustering algorithms are applied in order to discover semantic structures.⁴ Results of these analytical methods are compared to other approaches like Topic Modelling.⁵

Two independent corpora are used as separate test cases for the development and application of the methodological approach: The Buddhist Pali Canon and a selection of ancient Egyptian texts. These diverse corpora are used as a test bed for the development and application of language independent analytical tools. These corpora are available only in project specific formats. In order to have a generic starting point for analysis, they are converted to TEI compliant XML.

The Pali Canon is a huge collection of Buddhist texts in the middle Indic language Pali. It has been composed around the Common Era in Northern India and Sri Lanka and is nowadays available in a digital version.⁶ The Canon contains a lot of narratives in which the Buddhist authors depict the Buddha and his followers in relation to competing religious groups or individuals. Therefore the Pali Canon is a rich source for the analysis of interreligious dynamics in ancient South Asia. As the Canon is too huge for manual analysis, computational analysis is needed to get a clear picture of the social-semantic structures inherent in these narratives.

The database Thesaurus Linguae Aegyptiae⁷ represents an annotated selection of ancient Egyptian texts from different genres and periods. In our case it is used to analyse religious dynamics in the history of Ancient Egypt. An application of network analysis shall open up another perspective for an interpretation of interreligious contacts between Egypt and for instance the Near East. For this purpose the relations between actors, and between actors and concepts are extracted from syntactic and semantic structures.

The project contributes to the development of new analytical tools for the humanities. These should be of use for further research beyond the project. At the same time, significant results are expected for the fields of Egyptology, Indology, and the study of religions. Building on past digitisation efforts, the project allows for new insights based on the computational analysis of the corpora, complementing conventional hermeneutic approaches.

References

1. cf. Krech, Volkhard. (2012). *Religious Contacts in Past and Present Times: Aspects of a Research Programme*. Religion 42 (2): 191–213. doi:10.1080/0048721X.2012.642572.
2. cf. Gramsch, Robert. (2013). *Das Reich als Netzwerk der Fürsten: politische Strukturen unter dem Doppelkönigtum Friedrichs II. und Heinrichs (VII.)* 1225–1235.

3. cf. Biemann, Christian, Stefan Bordag, Gerhard Heyer, Uwe Quasthoff, and Christian Wolff. (2004). *Language-Independent Methods for Compiling Monolingual Lexical Data*. In Computational Linguistics and Intelligent Text Processing, edited by Alexander Gelbukh, 217–28. Lecture Notes in Computer Science 2945. Springer Berlin Heidelberg. link.springer.com/chapter/10.1007/978-3-540-24630-5_27.
- Ferrer i Cancho, Ramon, Ricard V. Solé, and Reinhard Köhler. 2004. "Patterns in Syntactic Dependency Networks." Physical Review E 69 (5): 051915. Ferrer2004. doi:10.1103/PhysRevE.69.051915. Paranyushkin, Dmitry. 2011. "Identifying the Pathways for Meaning Circulation Using Text Network Analysis". Nodus Labs. noduslabs.com/research/pathways-meaning-circulation-text-network-analysis/.
4. cf. Dorow, Beate, and Dominic Widdows. (2003). *Discovering Corpus-Specific Word Senses*. In Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics - Volume 2, 79–82. EACL '03. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.3115/1067737.1067753. dx.doi.org/10.3115/1067737.1067753.
5. cf. Blei, David M., Andrew Y. Ng, and Michael I. Jordan. (2003). *Latent Dirichlet Allocation*. J. Mach. Learn. Res. 3: 993–1022.
6. tipitaka.org
7. aaew.bbaw.de/tla

Taking manuscripts apart, and putting them together

Emery, R. Douglas

emery@upenn.edu

Schoenberg Institute for Manuscript Studies, University of Pennsylvania

Porter, Dot

dot.porter@gmail.com

Schoenberg Institute for Manuscript Studies, University of Pennsylvania

Campagnolo, Alberto

alberto.campagnolo@gmail.com

University of the Arts, London

Introduction

Our poster will demonstrate recent work on generating visualizations of codex manuscript structure, and visualizations of reconstructed historical manuscripts known through palimpsest. Next steps, presented at DH2014, will include linking the visualization to digital images of the manuscript pages, and work to parse more complex collation formulas. The purpose is to enable scholars and students to consider the manuscript, as a physical object, in ways not otherwise possible.

Taking manuscripts apart

We have started our collation visualization proof-of-concept work using the collation formulae from The Digital Walters TEI manuscript description files.¹ They use a simple subtraction-only system that describes the manuscript as it exists now, but makes no claims for how it came to be that way. This distinguishes it from other manuscript collation formulae, which typically distinguish between leaves added and leaves deleted.² In the Walters formula added folios are treated as half a sheet, of which the other half has been removed. For example, a manuscript with a leaf added between the second and third quires, and with a leaf removed in the middle of the fourth quire, would be described using this formula:

1(8), 2(10, -1), 3(8), 4(10, -6)

Quire 1 has eight leaves. Quire 2 has nine leaves, with the last leaf added (=the conjoin of the last leaf noted as missing).

Quire 3 has eight leaves. Quire 4 has nine leaves, with the sixth leaf missing.

Although it lacks the description found in other formulas developed for manuscripts and printed books, it has the benefit of being simple to parse - perfect for a proof-of-concept.

We find all of the collation formulas in the Digital Walters manuscript descriptions (not all of them include a formula), and then parse the formulas to generate a simple XML file describing the number of quires, number of leaves in each quire, and any "missing" leaves.

```
<quires>
  <quire n="1" leaves="8"/>
  <quire n="2" leaves="10">
    <less>1</less>
  </quire>
  <quire n="3" leaves="8"/>
  <quire n="4" leaves="10">
    <less>6</less>
  </quire>
</quires>
```

We then use this file to generate a set of SVG files, one for each quire. Leaves marked with <less> are currently indicated using outline, which existing leaves are indicated using dark lines.

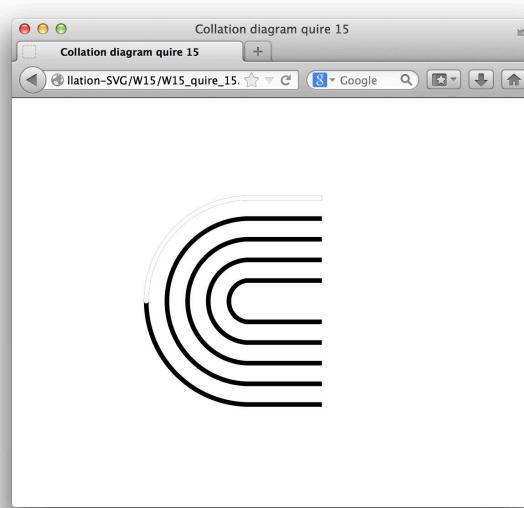


Fig. 1: Sample quire, missing the first leaf

At this point in the project (October 2013), we have the building blocks we need to begin constructing interactive collation visualizations for the manuscripts in The Digital Walters. Below, we will outline next steps for the project.

Putting manuscripts together

Our work with reconstructing palimpsest manuscripts is within the context of the Sinai Palimpsest Project and was initially presented at the TEI Member's Meeting and Conference, Rome 2013.³ The situation with the Sinai palimpsests is complex. Manuscripts were erased (the ink scraped off) and the pages re-used in other manuscripts. In some cases, fragments from several different historical manuscripts have been stitched together to create new leaves. Some fragments have been re-used more than once, resulting in layers of palimpsest.



Fig. 2: A sample folio from Sinai Arabic NF 8, consisting of four fragments from different historical manuscripts, stitched together to form a new folio.

We are developing a system using TEI to describe both historical manuscripts (the manuscripts from which the palimpsests originated) and the extant manuscripts (the manuscripts as they exist now, with parts of historical manuscripts distributed through them), and linking the two together while also creating visualizations of the historical manuscripts. We are interested not only in reconstructing individual palimpsest folios, but also in reconstructing complete manuscripts (as has already been done for the so-called Archimedes Palimpsest)⁴. This is the point at which our work taking manuscripts apart and our work putting manuscripts together overlaps: creating visualizations of the physical structure of both existing and historical manuscripts.

Poster session and next steps

At DH2014 we will present our most recent work on both taking manuscripts apart, and on putting them back together. Our poster will highlight the methods we use: how we describe our data, what we use to process the data, and visualization practices. We will also be prepared to demonstrate the methods in real time, particularly the process for translating a collation formula into a visualization.

Next steps will focus on expanding the collation visualization approach to more descriptive collation formulas⁵

, linking the collation visualization to digital images of the manuscript folios, and presenting the collation visualizations within an interface that would act as a workspace for scholars.

References

1. *The Digital Walters*, www.thedigitalwalters.org/Data/WaltersManuscripts/ManuscriptDescriptions/
2. Clemens, R. and T. Graham (2007). *Introduction to Manuscript Studies* (Cornell University Press), pp. 130-131

3. Emery, D. and D. Porter, (2013), "TEI and the Description of the Sinai Palimpsests" presented at The Linked TEI: Text Encoding in the Web, TEI Conference and Member's Meeting 2013, Rome, Italy. digilab2.let.uniroma1.it/teiconf2013/program/papers/abstracts-paper#C140

4. Netz, Reviel et al. *The Archimedes Palimpsest Volume 1: Catalogue and Commentary* (Cambridge University Press, 2011)

5. Bowers, Fredson. (1962). *Principles of Bibliographic Description*. New York: Russell & Russell.

Visualizing Homelessness

Engel, Maureen

mengel@ualberta.ca
University of Alberta

Zwicker, Heather

hwzicker@ualberta.ca
University of Alberta

Frizzera, Luciano

dosreisf@ualberta.ca
University of Alberta

Pedraça, Samia

pedraca@ualberta.ca
University of Alberta

Regattieri, Lorena

regattie@ualberta.ca
University of Alberta

Schoenberger, Zachary

zschoenb@ualberta.ca
University of Alberta

Windsor, Jennifer

jwindsor@ualberta.ca
University of Alberta

This poster presentation asks the question: How can data visualization techniques facilitate new analyses of homelessness?

The project originates with a partnership between the Edmonton Pipelines research group at the University of Alberta, Canada, and Homeward Trust, a “community-based, comprehensive housing organization that provides leadership and resources toward ending homelessness in Edmonton [Alberta].”¹ Since 1999, Homeward Trust has produced a bi-annual “Homeless Count” – a one-day snapshot of homelessness in the city. The 10 “Counts” produced to date, alongside the city of Calgary’s, represent the largest number of homeless counts conducted in any Canadian city. The data are collected manually (by volunteers), compiled into spreadsheets, and a hard copy research report is produced. Importantly, this report then becomes a key tool for policy makers, social service agencies, and all levels of government.

Homeward Trust approached the Pipelines team in 2012 to ask for assistance in producing alternate format reports, in order to more fully use and understand the data they were collecting. They were most immediately interested in mapping their data, which fits with the expertise of the Pipelines team, but over the course of our collaboration, they’ve become increasingly interested in how they can use other forms of visualizations to fulfill their mandate.

This poster is the result of our collaboration. It includes: 1) an outline of the methodology used by Homeward Trust to produce “The Count,” including their rationale for it its known limitations; 2) a number of experimental visualizations of the data, including heatmaps and infographics; 3) a series of arguments about what we can learn about homelessness based on these visualization techniques; and 4) some proposals for future research, including supplemental data gathering techniques and historical comparison, in order to more fully avail ourselves of the possibilities that the visualization

techniques make possible for community-based research and action.

To date, Homeward Trust has relied on conventional communication techniques to share its results. It produces a narrative report of approximately 40 pages that provides a narrative summary of the data, and a series of graphs and tables (bar graphs, pie charts etc.) that conventionally represent demographic data. The content of the graphs and tables are also conventional given the dataset: what percentage of homeless people are men? Are aboriginal? Are within certain age ranges? Certain attributes are also cross-referenced, so the count provides an account of, for example, gender broken down by age or by type of homelessness (but not both). Overall, it is a robust piece of work produced with very few resources, but the written report is the only public record of the research and the data.

The visualizations that our team has produced are able to ask very different questions and produce different results. By producing heatmaps, for example, we are able to ask whether particular age or ethnicity groups tend to gather in particular areas and not others. In identifying those areas, we can begin to hypothesize what particular affordances a specific location might offer to a particular demographic. Similarly, by geolocating the data and overlaying it with known social service agencies, we're able to analyze whether homeless people tend to congregate around those services designed for them. If so, which ones? Again this allows us to hypothesize about why that might be the case (or not, as the case may be). Given the extent of the dataset, we can also track these dynamics over time to see how these populations have moved over the course of more than a decade. We can even make inferences about whether particular forms of urban renewal and development push homeless populations out of their traditional comfort zones.

These serve as a few examples of how specifically geo-located visualizations can assist us in interpreting the data. Beyond the geo-spatial models, however, we are also producing infographics, both static and dynamic, as further means of facilitating access to the data. The static infographics are being purposefully designed as educational tools to be used by Homeward Trust in their outreach activities with the broader population. For example, the infographic that displays the basic demographic information deliberately uses human figures as icons in order to remind its audience of the humanity of the subject at hand; similarly, a second infographic asks its users to "how many homeless people are like me?" and allows users to input their own age, gender, ethnicity etc. and to see themselves reflected in the homeless population. Lastly, a graph database, currently under construction, will allow us to visualize many more relations and combinations than the static infographics are able to portray, and will facilitate exploration of the data by both service providers and the general public.

At the first stages of this project, we are limiting our approach to one year's data (2012). The next stage will be to broaden our reach to the historical data sets and to begin to formulate comparisons across time.

References

- Homeward Trust. 2012 Edmonton Homeless Count. www.homewardtrust.ca/images/resources/2013-01-22-11-53FINAL%20%202012%20Homeless%20Count.pdf

Digital Actors' Parts: An Interactive Tool for Learning Shakespeare's Plays

Estill, Laura

Texas A&M University, United States of America

Meneses, Luis

Texas A&M University, United States of America

In the early modern period, rather than having access to a full-text play, actors learned their lines using "Actors' parts," hastily handwritten documents that provided them with only their cues and lines. Traditionally, today's actors learn their lines from full-text plays, without any computer assistance. *Digital Actors' Parts (DAP)* is an online environment that both mimics and enhances the early modern acting experience in order to facilitate actors learning their lines. DAP is the first project to give users an interactive experience with an early-modern-inspired "actor's part," which encourages both active reading and memorization, in turn leading to a better understanding of the texts themselves.

In *Orality and Literacy: The Technologizing of the Word*, Walter Ong theorized that we were in a "second orality" based on the "use of writing and print" and considered how memorization worked in oral and written circumstances; *Digital Actor's Parts* allows us to consider memory as it relates to online technologies. In Shakespeare studies, the function of memory has been an important area of study for years, as scholars debate whether certain texts are "memorially reconstructed," that is, printed from actors' memories rather than a playwright's written text. Our tool argues for the importance of understanding memorization when it comes to cognition and the comprehension of literary and theatrical works. *DAP* will not only provide expert scholars with a platform for reconsidering memory specifically as it relates to Shakespeare, it will also offer a tool that will be of use for beginners and professionals both in the classroom and the theatre.

As the most prominent figure in English theatre, Shakespeare's plays have been encoded multiple times; digital Shakespeare projects abound. Unlike most single-source projects, *Digital Actors' Parts* brings together open-access data from multiple sources, including the encoded texts from the MLA Committee on the New Variorum Edition of Shakespeare, the Folger Shakespeare Library's Digital Texts, the Internet Shakespeare Editions, and Open Source Shakespeare. Our project is designed to incorporate new data from these sources, because, for instance, the MLA and Folger Digital Texts are still encoding Shakespeare's works and have not yet released their complete data-set. These online Shakespearean texts are as diverse as the original printed editions. By building on existing Shakespearean projects, *DAP* is not limited to a single copy-text, so an actor or director could choose either the folio or quarto version of, for instance, *King Lear*.

Digital Actors' Parts helps students, researchers, and theatre practitioners learn lines by allowing them to select a Shakespearean play, edition, and character from a pre-populated dropdown menu. The users are then presented with their first cue and the ability to type in their line(s). Users will also be able to ask for prompts: their first word, first line, or entire speech. By entering the text of their speeches and checking what they have typed, users receive a score depending on their accuracy. This feature makes Shakespeare's text interactive in a way beyond most digital editions. The accuracy score allows for the potential gamification of becoming a Shakespearean actor, which will especially promote student learning. In the classroom, *DAP* encourages students and teachers to memorize and perform Shakespeare's language, that is, to engage with it beyond simply reading. This online tool makes it easy to transition from, as the saying goes, the page to the stage.

Although Shakespeare's plays continue to be popular both online and in the theatre, most digital tools for the study of Shakespeare are aimed at literary scholars, teachers, and K-12 students. The intended audience of *Digital Actors' Parts* also includes theatre professionals (actors, directors, dramaturgs); indeed, this is not simply a digital humanities project, but also a digital fine arts endeavor.

Digital Actor's Parts is currently a work in progress. The prototype for *DAP* runs on an instance of the Web.py framework. Web.py is a Python web framework that is both simple and powerful. We chose to use this framework because it allows rapid prototyping allowing us to concentrate on the specific features that we need to implement. On the other hand,

the plays that we are utilizing are encoded and distributed in different formats. For example: The plays from Folger Digital Texts are encoded in TEI-compliant XML, MLA's are encoded only in XML, and the materials from Open Source Shakespeare are bundled in a Microsoft Access database. Therefore, before being displayed through the interface, the plays must be harvested from their original institutional repositories, parsed, transformed using XSL Transformations and stored into a database. The accuracy scores and metadata for the individual user records are maintained and stored in a using a separate table structure. Figure 1 shows a prototype of the user interface.

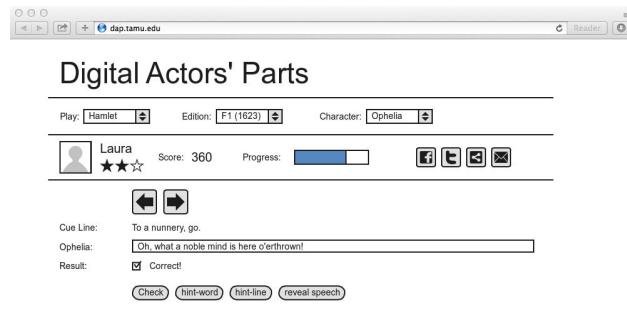


Fig. 1: Prototype of the user interface for Digital Actors' Parts.

Unlike other online memorization tools such as Memorizer and Memorize Now, this project does not require users to first input the text they memorize—which could be a particularly fraught process for texts as complicated as Shakespeare. *Digital Actors' Parts*, furthermore, is ideal to help users memorizing all of a character's lines (and not just a single speech) as it delivers the appropriate cues, unlike existing tools. The automatically-generated accuracy score, another feature other programs lack, can add elements of fun and competition as users strive for mastery of Shakespeare's text.

In its first iteration, *Digital Actors' Parts* relies on a modern web browser to deliver its content and experience. For future releases, we hope to implement an application for mobile devices. This app will allow users to expand the possibilities of interaction by taking DAP to the rehearsal space, theater, or experimenting with entirely new locations. Having this tool in hand will encourage earlier inclusion of blocking and business in the rehearsal process and will also be useful for active classroom learning.

In 2014, Shakespeare 450 will mark the anniversary of Shakespeare's birth with performances, workshops, and lectures. We envision *Digital Actors' Parts* participating in this global celebration by making it easier for performance troupes from amateur to professional to take part in staging Shakespeare's plays around the world. In "To the Memory of my Beloved the Author, Mr. William Shakespeare," Ben Jonson famously declared that Shakespeare was "not of an age, but for all time": DAP is part of the larger movement bringing Shakespeare into the twenty-first century with new digital resources and tool. DAP will help us understand how we engage with Shakespeare's works in the most fundamental ways and will allow us to theorize memory as it relates not simply to orality or written texts, but also to innovative, interactive, digital tools.

Ultimately, *Digital Actors' Parts* capitalizes on the proliferation of open-source Shakespeare texts, offering one answer to the question of "where do we go from here?" with digital projects. This project goes beyond aggregation by suggesting one way these texts can be fruitfully combined. While offering a valuable rehearsal tool in itself, DAP also encourages further research on Shakespeare's works, the digital practice of combining multiple corpora, and interactive online learning methods.

References

Simon Palfrey and Tiffany Stern, *Shakespeare in Parts*, Oxford: OUP, 2007.

Walter J. Ong, *Orality and Literacy: The Technologizing of the Word*, London & New York: Methuen, 1982, 136. See also esp. 57-68.

On "memorial reconstruction" and the function of memory in Shakespearean texts, see **Laurie Maguire**, *Shakespeare's Suspect Texts: The 'Bad' Quartos and their Contexts*, Cambridge: CUP, 1996 and **Paul Werstine**'s work, especially "A Century of 'Bad' Shakespeare Quartos" *Shakespeare Quarterly* 50 (1999): 310-33.

MLA New Variorum Shakespeare encoding: <https://github.com/mlaa/nvs-challenge>

Folger Digital Texts: www.folgerdigitaltexts.org

The Internet Shakespeare Editions: ise.uvic.ca

Open-Source Shakespeare: www.opensourceshakespeare.org

Web.py: webpy.org

Memorizer: www.memorizer.me

Memorize Now: www.memorizenow.com

Shakespeare 450: www.shakespeareanniversary.org

Ben Jonson, "To the Memory of my Beloved, the Author, Mr. William Shakespeare," in Shakespeare's first folio (1623). Facsimile: internetshakespeare.uvic.ca/Library/facsimile/book/SLNSW_F1/9

Combining Macro- and Microanalysis for Exploring the Construal of Scientific Disciplinarity

Fankhauser, Peter

fankhauser@ids-mannheim.de
IDS Mannheim, Germany

Kermes, Hannah

h.kermes@mx.uni-saarland.de
Saarland University, Germany

Teich, Elke

e.teich@mx.uni-saarland.de
Saarland University, Germany

1. Introduction

The English Scientific Text Corpus (SciTex) consists of about 5000 scientific papers with about 34 Mio tokens in two time slots, 1970/80s and 2000s^{1, 2}. It has been compiled to investigate the construal of scientific disciplinarity, in particular, how interdisciplinary contact disciplines emerge from their seed disciplines. Both time slots consist of nine disciplines: Computer Science (A) as one seed discipline, Linguistics (C1), Biology (C2), Mechanical Engineering (C3), Electrical Engineering (C4) as the other seed disciplines, and Computational Linguistics (B1), Bioinformatics (B2), Digital Construction (B3), and Microelectronics (B4) as the corresponding contact disciplines between A and C1-C4. The individual articles are subdivided into Abstract, Introduction, Main, and Conclusion.

The orthogonal dimensions time, discipline, and logical structure provide for many, potentially interesting setups of variational analysis: We can explore the diachronic evolution of contact disciplines in comparison to their seed disciplines, variation between contact disciplines and their seed disciplines, and genre variation between abstracts and text bodies in individual disciplines. In this paper we present an approach that combines a macroanalytic perspective³ with the more traditional microanalytic perspective served by concordance search to explore variation along these dimensions.

2. Approach

2.1. Macroanalysis

For supporting explorative macroanalysis, we use well understood visualization techniques – heatmaps and wordclouds – and combine them with intuitive exploration paradigms – drill down and side by side comparison (see Figure 1). The heatmaps and wordclouds are interactive, allowing for a closer inspection at various levels. The leftmost heatmap visualizes the highest level contrast between abstracts and text bodies in the two time slots (1970s/80s and 2000s). The middle and right heatmaps serve for inspecting a chosen contrast at a lower level at the level of individual disciplines. A particular contrast can be chosen by clicking on the respective square, numbers indicating which contrast is displayed in the middle (Selection 1) and right heatmap (Selection 2). In this example, the middle heatmap visualizes the distances between abstracts and text bodies, and the right heatmap visualizes the distances between text bodies and abstracts.

The wordclouds underneath the heatmaps display the most typical words for a chosen contrast. In Figure 1 the wordcloud to the left visualizes the most typical words for abstracts as opposed to text bodies in the 2000s. Unlike in the common use of wordclouds, the size of words is proportional to their contribution to the distance (as defined in Section 2.2), whereas relative frequency is visualized by color, ranging from purple to red.

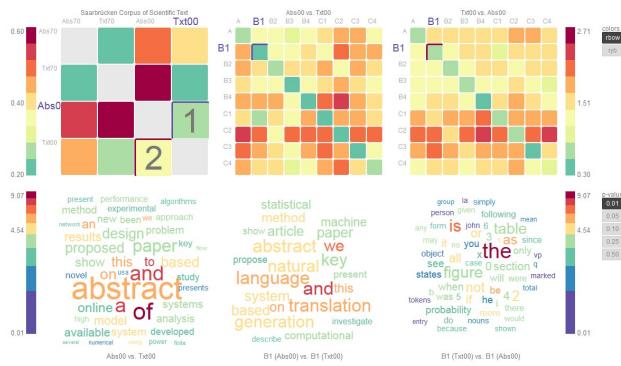


Fig. 1: Contrast between Abstracts and Text Bodies

Having a closer look at Figure 1, we can observe that the distance between abstracts is generally larger than the distance between text bodies, and that it has increased in the 30 years period. This general trend is mirrored in the individual disciplines (not shown here). Looking at the middle and right heatmaps, we can see that - not surprisingly - the distance between particular disciplines are at a minimum (squares forming the main diagonal), and the distances among the seed disciplines (A and C corpora), are generally larger than the distances among contract disciplines.

The corresponding wordclouds visualize the most typical words for abstracts (middle heatmap) and for text bodies (right heatmap) in the discipline B1 (Computational Linguistics). In this particular contrast, words typical for abstracts are clearly centered around constructions of exposition (*we propose, describe, investigate*), main topics of B1 (*natural, language (generation), machine translation*), words describing the methodology (*method, statistical, computational, system*) and function words (*and, of, on*). Words typical for text bodies are markedly different: they comprise B1's main entities of topic elaboration (*tokens, nouns, object, vp, john, probability*), references (*see figure, table, section*), conjunctions (*when, since, because, if*), auxiliary and modal verbs (*be, is, was, were, do, will, would, may*), and prominently, the determiner *the*. In summary, abstracts exhibit characteristics of an informationally dense text (e.g., omission of determiners) with topic oriented content. In contrast, text bodies are less dense (determiners, references, modality) and more elaborated.

Other contrastive pairs, such as the synchronic comparison between disciplines or the diachronic comparison of the two time slots, corroborate the results derived by means of computationally much more demanding techniques from machine learning [1], [2].

2.2. Corpus Representation and Distance Measures

The individual corpora are tokenized, and tokens are transformed to lower case. Stopwords are deliberately not excluded to inspect all levels of variation: style, lexico-grammar, and theme. On this basis, corpora are represented by means of unigram language models smoothed with Jelinek-Mercer smoothing, which is a linear interpolation between the relative frequency of a word in a subcorpus and its relative frequency in the entire corpus⁴. The distance between two corpora P and Q is measured by relative entropy D , also known as Kullback-Leibler Divergence:

$$D(P||Q) = \sum_w p(w) \log_2(p(w)/q(w))$$

Here $p(w)$ is the probability of a word w in P , and $q(w)$ is its probability in Q . Relative entropy thus measures the average amount of *additional* bits per word needed to encode words distributed according to P by using an encoding optimized for Q . Note that this measure is asymmetric, i.e., $D(P||Q) \neq D(Q||P)$, and has its minimum at 0 for $P = Q$ ⁵.

The individual word weights are calculated by the pointwise Kullback-Leibler Divergence⁶:

$$D_w(P||Q) = p(w) * \log_2(p(w)/q(w))$$

For all words the statistical significance of a difference is calculated based on an unpaired Welch t-test on the observed word probabilities in the individual documents of a corpus. This is used for discarding words below a given level of significance (p-value). A more detailed comparison with other measures for comparing corpora⁷ is beyond the scope of this paper and will appear in another venue.

2.3. Microanalysis

Wordclouds serve as a bridge between the big distance picture of macroanalysis and microanalysis. To this end, they are seamlessly integrated with the IMS Open Corpus Workbench (CQPWeb: <http://cwb.sourceforge.net/index.php>), which provides for an expressive corpus query language and several summarization tools, such as collocations and comparative word frequency lists. A click on a word sends a query to CQPWeb, which returns the word in the chosen context. For example, clicking on “do” in the right heatmap (B1 (Txt00) vs. B1 (Abs00)) generates the following query shown in Figure 2.

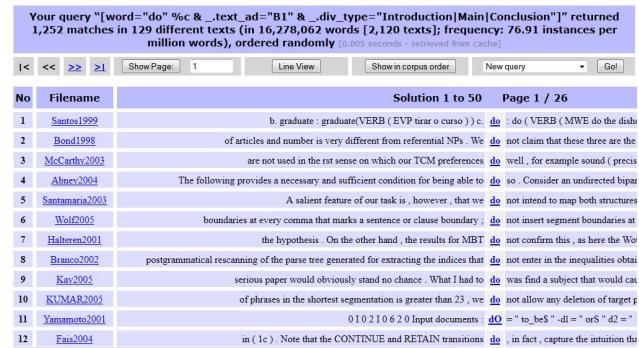


Fig. 2: Concordance for “do” in B1, text bodies, 2000s

This query returns a concordance for “do” in the 2000s slot of SciTex constrained to subcorpus B1 and to the divisions Introduction, Main, and Conclusion. Based on this list one can inspect the larger context of individual hits and get a ranked list of collocations to distinguish the uses of “do” as an auxiliary vs. main verb.

3. Related Work

The need for combining macroanalysis with microanalysis is well recognized in the DH community^{8, 9}, and there does exist a variety of frameworks with similar goals. Due to space restrictions, we can only give an exemplary selection; for a comprehensive overview see TAPoR 2.0 (<http://tapor.ca/>).

The MONK workbench¹⁰ allows to compare pairs of corpora using Dunning's log-likelihood ratio¹¹ for word weighting. Apart from the different distance measure, the main difference of our approach is that we combine the macro perspective of overall distance with the micro perspective of individual word weights to allow for an explorative analysis of variation. The Voyant Tools¹² provide a plethora of text visualizations, including word clouds, cooccurrences, and word trends based on frequencies. The focus of these tools, however, lies on summarizing and visualizing one text or corpus, rather than on exploring variation among corpora. Finally, the TXM platform¹³ integrates the IMS Corpus Workbench with some macroanalysis R packages such as factorial correspondence analysis, contrastive word specificity, and cooccurrence analysis. While this integration certainly provides a broader set of analysis techniques, it is arguably more complicated to use than the system presented in this paper.

4. Summary and Future Work

We have presented a system that combines macroanalysis with microanalysis to explore language variation, and briefly illustrated its use for analyzing differences along the dimensions time, discipline, and genre in a corpus of scientific text. Future work will be devoted both to technical as well as methodological enhancements. A useful technical extension is the facility to interactively group subcorpora to larger units, maybe with the help of hierarchical clustering based on the distance matrix to form meaningful groups. More generally, the support for importing external corpora and exporting distance matrices and word weights for analysis with other tools is desirable – the presented system has been evaluated based on a number of corpora, but the underlying processing pipeline certainly needs to be generalized and improved. On the methodological side the main challenge lies in supporting a broader variety of feature sets beyond simple unigram language models. This includes latent language models such as topic models¹⁴ and hidden markov models¹⁵, but also enriched representations such as part-of-speech tagging, and other extensions of unigram models. Such richer feature sets allow to focus analysis by means of feature selection, but also bear new challenges in measuring and visualizing the contribution of features to a contrast at hand, and translating features into meaningful queries against the underlying corpus.

References

1. **Elke Teich and Peter Fankhauser** (2010). *Exploring a Corpus of Scientific Texts using Data Mining*. In S. Gries, S. Wulff, and M. Davies, editors, *Corpus-linguistic applications: Current studies, new directions*, pp. 233–247. Rodopi, Amsterdam and New York.
2. **Stefania Degaetano-Ortlieb, Hannah Kermes, Ekaterina Lapshinova-Koltunski, and Elke Teich** (2013). *SciTex: A diachronic corpus for analyzing the development of scientific registers*. In Paul Bennett, Martin Durrell, Silke Scheible, and Richard J. Whitt, editors, *New Methods in Historical Corpus Linguistics, Corpus Linguistics and Interdisciplinary Perspectives on Language (CLIP)*, Volume 3, Narr, Tübingen.
3. **Matthew L. Jockers** (2013). *Macroanalysis: Digital Methods & Literary History*. University of Illinois Press, Urbana, Chicago, and Springfield.
4. **Chengxiang Zhai and John Lafferty** (2004). *A study of smoothing methods for language models applied to information retrieval*. ACM Transactions on Information Systems (TOIS), 22(2):179–214.
5. **David J. C. MacKay** (2002). *Information Theory, Inference & Learning Algorithms*. Cambridge University Press, New York, NY, USA.
6. **Takashi Tomokiyo and Matthew Hurst** (2003). *A language model approach to keyphrase extraction*. Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment (MWE '03), Vol. 18, Association for Computational Linguistics, Stroudsburg,

PA, USA, pp. 33–40. DOI=10.3115/1119282.1119287
dx.doi.org/10.3115/1119282.1119287

7. **Adam Kilgarriff** (2001). *Comparing Corpora*. International Journal of Corpus Linguistics, 6(1):97–133.
8. **Michael Correll and Michael Gleicher** (2012). *What Shakespeare Taught Us About Text Visualization*. IEEE Visualization Workshop Proceedings, 2nd Workshop on Interactive Visual Text Analytics: Task-Driven Analysis of Social Media Content, Seattle, Washington, USA, Oct 2012.
9. **Matthew L. Jockers and Julia Flanders** (2013). *A Matter of Scale*. Staged debate at the Boston Area Days of Digital Humanities Conference at Northeastern University, March 18, 2013. digitalcommons.unl.edu/englishfacpubs/106/
10. **John Unsworth and Martin Mueller** (2009). *The MONK Project Final Report*. Sep 2009. www.monkproject.org/MONKProjectFinalReport.pdf
11. **Ted Dunning** (1993). *Accurate methods for the statistics of surprise and coincidence*. Computational Linguistics 19(1):61–74.
12. **Stéfan Sinclair, Geoffrey Rockwell and the Voyant Tools Team** (2012). *Voyant Tools* (web application). http://voyant-tools.org/
13. **Serge Heiden** (2010). *The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme*. Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation, Institute for Digital Enhancement of Cognitive Development, Waseda University, Japan, Nov 2010, pp. 389–398.
14. **David. M. Blei, Andrew Y. Ng, and Michael I. Jordan** (2003). *Latent Dirichlet Allocation*. Journal of Machine Learning Research, 3:993–1022.
15. **Sharon Goldwater and Tom Griffiths** (2007). *A fully Bayesian approach to unsupervised part-of-speech tagging*. Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL'07). Association for Computational Linguistics, Prague, Czech Republic, June 2007, pp. 744–751. www.aclweb.org/anthology/P07-1094

Reading Again: Annotating, Editing, and Writing in the Browser. Pedagogy, Design, and Development of Annotation Studio

Fendt, Kurt

fendt@mit.edu
MIT, Comparative Media Studies/Writing

Kelley, Wyn

wkelley@mit.edu
MIT, Literature

Folsom, Jamie

jfolsom@mit.edu
MIT, Comparative Media Studies/Writing

1. Keywords

Annotation, collaboration, pedagogy, critical thinking, writing, close reading, software methodology

2. Pedagogy

“By re-reading and looking for meaning in the text, you start to think critically.”

~ Student, First-Year Writing, Hofstra University, using Annotation Studio

“Within the humanities,” argue Anne Burdick, Johanna Drucker, et al in their 2012 Digital Humanities, “close reading has been a central practice that is premised on careful attention to features contained in a text” (our emphasis)¹ Advances in

the digital humanities have “opened up new ways of creating meaning through distant reading” (our emphasis), and scholars have debated the significance of each, but the authors of Digital Humanities emphasize a dynamic relation between the two: “Rather than pitting distant reading against close reading, what we are seeing is the emergence of new conjunctions between the macro and the micro, general surface trends and deep hermeneutic inquiry, the global view from above and the local view on the ground” (39). In developing Annotation Studio, a digital pedagogical tool for selecting, commenting on, and sharing texts within a humanities classroom, the research team at HyperStudio, the Digital Humanities Lab at the Massachusetts Institute of Technology, have built a simple, practical utensil that enables both close study and distant analysis. We first presented the tool at its early stages at the 2012 Digital Humanities meeting after we had just received a Start-Up grant from the National Endowment for the Humanities (NEH). In this presentation we will discuss the results from the first development phase of Annotation Studio, including classroom integration, assessment research, and iterative software development process now that Annotation Studio has been tested extensively in a range of humanities classes at MIT and elsewhere. We will conclude with future directions for the tool that will be implemented over the next two years, supported by a NEH Digital Humanities Implementation Grant.

Allowing students to engage deeply with a text, Annotation Studio also expands their awareness of their own critical reading, writing, and thinking as taking place in a social, shared space. Social reading, we have found, lends itself to an enhanced experience of writing, as students recognize in the audience for their essays the very people with whom they have read their texts. In this fluid progress from reading to writing, we have seen something new emerge: a distinct sense among students of themselves as editors, as people who manipulate text from the earliest stages of the critical process. This sense of command of a text empowers students to read and write more confidently and with greater pleasure than they have come to expect in a typical humanities classroom.

Although useful in a number of scholarly applications, Annotation Studio shows particular promise for pedagogy, as it enables students to read and read again by selecting text, writing comments, sharing responses with other students, and tagging and sorting annotations for further research and writing. The tool builds on the deep history of marginalia of all kinds, from illuminated texts to private authorial markings to marginal text meant for coterie circulation to the comment features of contemporary word processing.² Adapting the history of marginalia to the classroom, Annotation Studio makes visible the earliest stirrings of critical thought and leaves a record of the reading process for both writers and readers to return to and absorb.

The pedagogy growing out of Annotation Studio also responds to new theories of media literacy.³ Drawing from Henry Jenkins' research on online fan communities, we have experimented with the use of shared annotation as a way for students to interact critically and creatively with texts and with each other. Recognizing that authors from the past borrowed from other writers and shared their texts, we have created a digital workspace where students can participate in a fluid relationship with texts from the past. At the same time, given an environment in which such fluid relationships can raise serious questions about intellectual property and theft, we use the digital text as an opportunity for discussion of responsible research, citation, and attribution. Annotation Studio, with its precise marking tools, allows students to see the boundaries between text and margin, between what is theirs and what comes from someone else's writing.

The use of annotation in a fluid-text environment lends itself as well to critical writing and especially to a sense of the student writer as an editor.⁴ If the first stage of the writing process involves meticulous marking of a text, the reader is already performing some of the functions of an editor: defining unfamiliar words, glossing historical references and literary allusions, hazarding analysis. When it comes time to write, the student has already practiced primary marking, research, and interpretation with an audience of peers, the other members

of the class. We have found that although students often find critical writing intimidating, they tend to feel less overwhelmed when they have experienced these initial steps and have come to think of themselves as editors more than as writers. Working in Annotation Studio, they get to raise the questions and work out the problems that make a text seem mystifying. Their papers improve as a result of this work, and their imaginations soar.

Since presenting Annotation Studio in 2012, we have developed the pedagogy considerably, as noted above. We have also greatly developed the tool, advanced the partnerships with instructors and developers that sustain it, and performed the assessment necessary for moving ahead.

3. Software Development

As we plan and implement new features in the Annotation Studio web application, we are careful to do so in response to the needs, preferences and demands of instructors and students in humanities classrooms at MIT and at partner institutions. The main features of the tool are all drawn from what those audiences have articulated as functionality they need in their disciplines.

Guided by that approach, we have written Annotation Studio using an agile software development methodology, which accommodates continuous feedback from users and permits refinements based on that feedback.

In addition, we have integrated an open source annotation engine from the Open Knowledge Foundation, called the Annotator, which has an open architecture that has made it possible for us to add functionality specific for the educational context. Further, we have released Annotation Studio itself under an open source license. Engaging in this way with other communities of developers has accelerated progress on our own software, and has made it possible to connect with coders and instructors at other institutions.

The combination of an iterative development process and engagement with other developers has produced an increasingly functional tool, and one which is also increasingly focused on the peculiar needs of the humanities classroom.

In that context, we will describe some architectural and user-facing features of Annotation Studio which will serve to illustrate our approach, including aspects of technical planning and decision-making, our process for selecting, sequencing and implementing features, and our experience collaborating on open source projects as those may be useful to others writing software for Digital Humanities.

4. Conclusion

With its tight integration into classroom practices and use of iterative software development practices, Annotation Studio has proven to be an easy to use yet powerful tool that has helped students practice traditional humanistic skills within a social environment. At the same time, the web application has made the reading process more transparent, not only to the instructor but to the students themselves. Such a renewed focus on close reading opens up exciting new possibilities for engaged forms of writing that have the potential to reflect source texts more deeply.

References

1. Anne Burdick, Johanna Drucker, Peter Lunenfeld, Todd Presner, Jeffrey Schnapp (2012), *Digital Humanities* Cambridge, MA: The MIT Press
2. See H. J. Jackson (2001), *Marginalia: Readers Writing in Books* New Haven, CT: Yale University Press.
3. Henry Jenkins, Wyn Kelley, Katie Clinton, Jenna McWilliams, Ricardo Pitts-Wiley, Erin Reilly (2013), *Reading in a Participatory Culture: Remixing Moby-Dick in the English Classroom* (New York: Teachers College Press).

4. On fluid texts and theories of editing, see **John Bryant**, *The Fluid Text: A Theory of Revision and Editing for Book and Screen* Ann Arbor, MI: University of Michigan Press, 2002.

Rethinking HathiTrust Metadata to Support Workset Creation for Scholarly Analysis

Fenlon, Katrina

kfenlon2@illinois.edu

University of Illinois at Urbana-Champaign, USA

Cole, Timothy

t-cole3@illinois.edu

University of Illinois at Urbana-Champaign, USA

Han, Myung-Ja

mhan3@illinois.edu

University of Illinois at Urbana-Champaign, USA

Willis, Craig

willis8@illinois.edu

University of Illinois at Urbana-Champaign, USA

Fallaw, Colleen

mfall3@illinois.edu

University of Illinois at Urbana-Champaign, USA

1. Introduction

The HathiTrust Digital Library includes over 10 million volumes digitized from more than 90 research libraries. The HathiTrust Research Center (HTRC) has been established to help scholars get the most from this massive text corpus by providing cutting-edge tools, services and cyberinfrastructure that enable advanced computational access to the HathiTrust corpus. An immediate objective for HTRC is to allow scholars to collect items together for computational analysis. This has required rethinking the HathiTrust metadata model, inherited from print-based library cataloging traditions. This poster describes the motivation for this work, shortcomings of the current metadata model, and requirements driving the updated model.

2. Motivation

Humanities scholars regularly create collections in the course of their research – selecting, gathering, and organizing materials from disparate sources to answer specific research questions^{1 2}. As scholars increasingly rely on digital sources, they need sophisticated tools for the management and manipulation of “custom collections” of digital texts^{3 4 5 6 7 8 9}.

The HTRC workset creation tools will allow users to formally gather selected subsets of the HathiTrust corpus together for computational analysis. Early user studies¹⁰ suggest several requirements, e.g.:

- Worksets must allow scholars to gather not just the primary constituents of the HathiTrust corpus (books), but also metadata and granular, intra-book content.
- Worksets must allow integration of external sources, such as linked datasets, secondary literature, and references, as shown in Figure 1.
- Scholars must be able to identify and describe worksets so that they may function as sustainable and reusable scholarly resources.

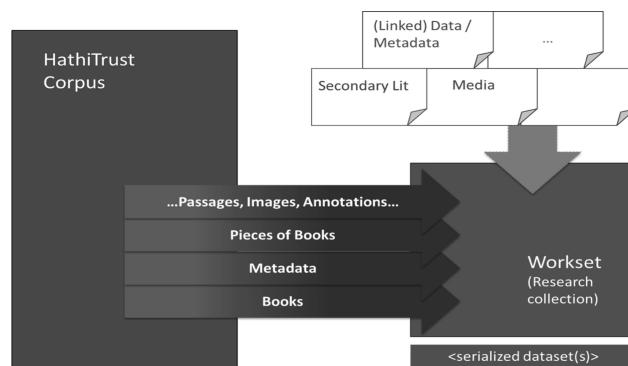


Fig. 1: Fig. 1: Creating worksets for scholarly analysis

3. Limitations of MARC-based metadata

Items in the HathiTrust corpus today are described exclusively by MARC. While MARC is the predominant bibliographic metadata standard used in libraries, it is proving inadequate to support the creation of scholarly worksets from large digital repositories such as the HathiTrust.

To begin with MARC can accommodate only a fraction of properties of texts and their contexts that are of interest to scholars. For example, the MARC bibliographic format does not provide fields for describing an author's gender, nationality, religion or social relationships. In addition, library catalogers rarely use the full expressiveness potential of MARC. The MARC specification defines more than 1,900 fields. However, most bibliographic records contain only a handful of these¹¹. Table 1 illustrates the use of MARC fields across the 6 million HathiTrust bibliographic records. Additionally fields used vary by class of text. Table 2 illustrates how infrequently subject headings are used in describing fictional works.

Property	Percent of Records Having
Title	> 99%
Publisher	87%
Subject -- Topical	72%
LC / Dewey Classification Number	41% / 17%
Subject -- Geographic	37%
Subject -- Temporal	10%
Fiction Literary Form	5%

Property	Percent of Fiction Records Having
Subject -- Topical	25%
Subject -- Geographic	10%
Subject -- Temporal	5%

4. Metadata Design Requirements to Support Workset Creation

With the generous support of the Andrew W. Mellon Foundation, the Workset Creation for Scholarly Analysis (WCSA) project, a collaboration between the HTRC and 4 independent research groups (competitively selected from among 15 respondents to a Request For Proposals issued in November 2013), is exploring answers to the following intertwined questions:

- Given sparseness of HathiTrust records, can we enrich the corpus metadata by distilling analytics over full text? Could we deploy/modify off-the-shelf tools, for example, to confirm or determine language(s) of the text, temporal coverage, spatial coverage, etc.?
- Can we augment string-based metadata with URIs for entities – e.g., names, subjects, place of publication, etc.? If so,

- HTRC could leverage additional services to meet scholars' needs.
- Can we formalize the notion of worksets in HTRC, e.g., defining the necessary elements of a workset? In doing so, how do we balance rigor with extensibility and flexibility? What roles do "data", "metadata", "annotations", "tags", "feature sets", and so on, play in the conception, creation, use and reuse of worksets?

In reporting on these questions, we expect to articulate recommendations to move away from a solely MARC-based metadata architecture towards a more RDF-centric metadata architecture relying on multiple library-specific and non-library standards, e.g., MARC, MODS, DC, SKOS, FOAF, schema.org, etc.

References

1. **Brogan, M.** (2006). *Contexts and Contributions: Building the Distributed Library*. Digital Library Federation/Council on Library and Information Resources. Retrieved August2, 2010 from www.diglib.org/pubs/dlf106
2. **Palmer, C. L.** (2005). *Scholarly work and the shaping of digital access*. Journal of the American Society for Information Science and Technology, 56(11), 1140-1153.
3. **Dempsey, L.** (2006). *The (digital) library environment: Ten years after*. Ariadne, 46. Retrieved February 13, 2013 from www.ariadne.ac.uk/issue46/dempsey/
4. **Green, H., Saylor, N., & Courtney, A.** (2013). *Beyond the scanned image: A needs assessment of faculty users of digital collections*. Digital Humanities 2013, Lincoln, Nebraska.
5. **Mueller, M.** (2010). *Towards a Digital Carrel: A Report about Corpus Query Tools*, retrieved September 17, 2013 from panini.northwestern.edu/mmueller/corpusquerytools.pdf
6. **Spiro, L., & Segal, J.** (2005). *The Impact of Digital Resources on Humanities Research*, retrieved October 31, 2013 from library.rice.edu/services/dmc/about/projects/the-impact-of-digital-resources-on-humanities-research
7. **Warwick, C., Terras, M., Huntington, P., & Pappa, N.** (2008). *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*. Literary and Linguistic Computing, 23(1), 85-102.
8. **Sukovic, S.** (2008). *Convergent flows: Humanities scholars and their interactions with electronic texts*. Library Quarterly 78(3), 263-284.
9. **Sukovic, S.** (2011). *E-Texts in Research Projects in the Humanities*. In A. Woodsworth & W. D. Penniman (Eds.) *Advances in Librarianship* (131-202). Bingley, UK: Emerald Group Publishing.
10. **Green, H., Fenlon, K., Senseney, M., Bhattacharyya, S., Willis, C., Organisciak, P., Downie, J. S., Cole, T., and Plale, E.** (2014). *Using collections and worksets in large-scale corpora: Preliminary findings from the Workset Creation for Scholarly Analysis project*. Forthcoming paper to be presented at iConference 2014, Berlin, Germany.
11. **Moen, William E. & Bernardino, P.** (2003). *Assessing Metadata Utilization: An Analysis of MARC Content Designation Use*. 2003 Dublin Core Conference: Supporting Communities of Discourse and Practice — Metadata Research and Application, Seattle, Wash. Retrieved October 31, 2013 from www.unt.edu/wmoen/publications/MARCPaper_Final2003.pdf

Enhancing Scholarly Communication and Communities with the PressForward Plugin

Fragaszy Troyano, Joan

joanfragaszytroyano@gmail.com

George Mason University, Roy Rosenzweig Center for History and New Media, United States of America

Rhody, Lisa

lmrhody@gmail.com

George Mason University, Roy Rosenzweig Center for History and New Media, United States of America

Coble, Zach

coblezc@gmail.com

New York University

Shirazi, Roxanne

roxanneshirazi@gmail.com

CUNY Graduate Center

Potvin, Sarah

sarah.potvin@gmail.com

Texas A & M

Pinto, Caro

pinto.caro@gmail.com

Mount Holyoke

The PressForward Initiative¹ at the Roy Rosenzweig Center for History and New Media (RRCHNM) has developed a methodology and a technology to expedite scholarly communications and nurture and expand communities of practice on the open web. Over the past two years we have produced our own open source WordPress plugin² to facilitate aggregating, curating, and disseminating scholarly content with a process that fosters community and resource-sharing among its users. Simultaneously, we experiment with multiple processes for surfacing, selecting, and circulating digital humanities work and gray literature outside formal traditional publication models.

This poster profiles the workflows and content of two digital humanities publications that have adapted technology developed by PressForward to suit their unique editorial and community needs: *Digital Humanities Now*³, produced by RRCHNM and *dh+lib*⁴, a publication hosted by the Association of College and Research Libraries' Digital Humanities Discussion Group. Visualizing and documenting intellectual and technical schemas for the plugins with diagrams and onsite demonstrations, this presentation exposes the philosophies and methodologies behind aggregating and curating scholarly work and learned expertise made available on the open web. Illustrating multiple workflows, layouts, and interfaces, this poster presents the scalable, replicable, and adaptable potential of the PressForward Plugin for niche scholarly communities eager to tailor their own hubs for communication and collaboration.

As a technology, the PressForward Plugin provides a smoothly integrated editorial process for the aggregation, review, discussion, and republication of external web content within the WordPress dashboard. PressForward aggregates content via RSS feeds, functions as a seamless feed reader, and allows users and groups to mark and discuss individual items before modifying or reproducing them for republication. This poster will document how two publications currently tailor the scope, structure, and flow of the plugin, gesturing toward the potential for replication and modification.

The largest case study, *Digital Humanities Now* (DHNow), is a principle test case for the plugin and for streamlining workflows that facilitate volunteers' nomination of scholarship on the open web. Drawing from a growing digital humanities community, volunteer editors-at-large sign up to survey over 1,000 potential items each week from more than 650 RSS feeds and nominate material they believe is salient to DHNow's readership. Nominations are considered for republication in DHNow as either Editor's Choice or News items. Next, rotating editors-in-chief -- faculty and graduate assistants at RRCHNM -- select, prepare, and publish links to nominated content, improving its visibility by directing attention back to the original site of publication. DHNow's streamlined processes for managing large numbers of novice and seasoned community volunteers also experiments with automated methods that include computer learning algorithms for filtering through large amounts of content.

dh+lib, published by an existing professional group, aims to give increased presence and voice to librarians interested in, or already knowledgeable about, DH. *dh+lib* publishes content in two streams. First, original content--posts, essays, and other

work--is published directly to the site biweekly or monthly. Second, more regular, content, appears as the *dh+lib Review*. To produce *dh+lib Review*, *dh+lib* relies on the PressForward plugin to facilitate the nomination of content from approximately 167 RSS feeds. Modeled after *DHNow*, *dh+lib*'s workflow also relies on volunteer editors-at-large, who sign up for weekly shifts to sift through the feeds to select material of interest to the community. *dh+lib* has also developed an additional layer of editorial intervention, where items selected for publication are written up as short "review" pieces. These pieces contextualize the nominated content, often pulling in other sources, links, and commentary to frame a project, resource, or post.

By volunteering for *DHNow* or *dh+lib*, community editors-at-large develop conversance in relevant trends and issues. While *DHNow* broadcasts a wide scope of work and provides an easy entry point for new practitioners to begin tracking the field, the published *dh+lib Review* provides librarians with weekly updates of useful and timely content, as selected by their peers. Additional examples will include *Global Perspectives on Digital History*⁵, a collaboration between RRCHNM and his.net, that has adapted the PressForward methodology to monitor and distribute material from a smaller, but multilingual source base.

Committed to the ongoing experimentation of *DHNow* and *GPDH*, PressForward also encourages adaptation of the tool by others and welcomes collaborations such as with *dh+lib*. Exposing the production of these digital humanities publications through documentation, workflow diagrams, and guides for getting started, we encourage viewers to consider how the PressForward model might improve the scholarly communication and collaboration of their own communities of interest on the open web.

References

1. *PressForward: Discover, Curate, and Distribute Scholarship the Web Way*. PressForward. Roy Rosenzweig Center for History and New Media, n.d. Web. 30 Oct. 2013. pressforward.org
2. *The PressForward Plugin*. PressForward. Roy Rosenzweig Center for History and New Media, n.d. Web. 30 Oct. 2013. pressforward.org/the-pressforward-plugin/
3. *Digital Humanities Now: Digital Humanities Now*. Roy Rosenzweig Center for History and New Media, n.d. Web. 01 Nov. 2013. digitalhumanitiesnow.org
4. *dh+lib*. Association of College and Research Libraries, n.d. Web. 01 Nov. 2013. acrl.ala.org/dh/
5. *Global Perspectives on Digital History*. N.p., n.d. Web. 01 Nov. 2013. gpdh.org

Check! – An online tool for the recognition and evaluation of DH work

Galina Russell, Isabel

Universidad Nacional Autónoma de México - UNAM, Mexico

Priani Saisó, Ernesto

Universidad Nacional Autónoma de México - UNAM, Mexico

Introduction

There has been an increase in the recognition of Digital Humanities (DH) projects as legitimate forms of Humanities research (MLA 2007). Although much work has been done to convince our colleagues that digital research is worthy, a pending problem is that they do not necessarily have the knowledge or tools to evaluate DH work (Rockwell, 2011). Most DH projects are not presented in traditional research output formats such as articles or books, but take on various formats such as development of metadata, textual markup, tools, websites and others (Schreibman et.al, 2011), so many

evaluation committees are at loss with how to deal with these materials. At the same time, people developing a DH project do not necessarily know what elements should be included.

Although there are places in the world where advanced and sophisticated DH projects exist, the vast majority are individual and modest efforts with little access to or knowledge about best DH practices. This issue becomes particularly acute in emergent DH communities, such as Mexico, where there is little accumulated practical experience. In previously held workshops of the RedHD (Galina 2012; Galina & Priani 2011), where issues related to developing and promoting DH in the region were addressed, we concluded that one important aspect was providing both DH creators as well as evaluators with mechanisms for recognizing and evaluating the importance and impact of DH projects.

We therefore decided to develop an online tool for the evaluation of DH project in order to promote the recognition and evaluation of DH work, consisting of a set of guidelines coupled with an online checklist that allows users to evaluate a particular resource using an interface that displays the results of the evaluation, indicating areas of weakness and of strength. The aim of this tool is threefold: step-by-step evaluation tool for committees, a resource for developers and as an informal compliance of a minimum standard.

Methodology

A literature review of relevant guidelines was undertaken and an ad hoc committee was created to discuss the findings. Other tools were found to be either too specific or general (MLA 2012; Presner 2012; MLA 2011; Warwick 2007; Unsworth 2001) and additionally they were all provided as static document-based resources.

Based on this, guidelines were proposed and divided into sections: Authorship and Attribution, Documentation, Quality Control, Rights Management and Visibility and Dissemination, were created. We then extracted simple yes, no or not applicable questions in order to build up a checklist that corresponded to the guidelines. The checklist was then transferred online and we developed a punctuation system depending on the answers. The results are then displayed by section and overall performance.

The tool was reviewed by an expert group of 16 people. Each person was assigned five DH projects to evaluate from a random sample taken from the RedHD DH project database. A follow up meeting with all participants was used to discuss their experience and several modifications were made to the tool until a consensus was reached.

Conclusions

Creating an online resource for evaluating and building a DH project was a necessary but challenging endeavor. DH resources are heterogeneous and their objectives vary widely. It is therefore difficult to condense into a checklist DH desirable characteristics that are not too specific but at the same time manage to incorporate issues that are particular to DH projects and not to web-based projects in general. In addition we wanted our tool to address different types of audiences: creators and evaluators of DH resources whom would have different degrees of DH knowledge and expertise. Choosing simple and clear language for the best practices was a challenge. In addition, as we were able to ascertain from the discussions during the committee meetings, there are no clear and definite answers and this tool will probably have to be adjusted continuously as DH work and the technologies evolve and will require continuous community consensus. In the following months the tool will be tried out with two other user groups: a group of inexpert colleagues who are developing for a DH project for the first time and an evaluation committee that provides small funding grants for UNAM projects. The feedback from these exercises will also be incorporated as the work on this tool continues. We do believe however, that this tool that provides an informal certification of compliance will aid in the process of recognition of the validity of DH projects as it gives

both creators and evaluators a general standard to which DH projects can be measured against.

References

- Galina, I.** *Retos para la creación de recursos digitales en las Humanidades*, El Profesional de la Información, 21(2), pp185-189, 2012 (ISSN: 1386-6710)
- Galina, I., Priani, E., López, J., Rivera, E., Cruz, A.** *Tejiendo la Red HD- A case study of building a DH network in Mexico*, Digital Humanities 2012, Conference abstracts, Hamburg, Germany. 16 - 22 July 2012, ISBN 978-3-937816-99-9, pp.456-458
- MLA** (2007). *Report of the MLA Task Force on Evaluating Scholarship for Tenure and Promotion*, 2007, Modern Language Association, p.71
- MLA** (2011), *Guidelines for Editors of Scholarly Editions*, Modern Language Association, 2011.
- MLA** (2012), Guidelines for Evaluating Work in Digital Humanities and Digital Media, Created 2000, last reviewed 2012.
- Presner, T.** (2011). How to evaluate digital scholarship, http://humanitiesblast.com/Evaluating_digital_scholarship.pdf
- Rockwell, G.** (2011) On the Evaluation of Digital Media as Scholarship, *Profession*, 2011, pp. 152-168, 10.1632/prof.2011.2011.1.152
- Schreibman, L., Mandell, L., et. al.**, (2011) Evaluating Digital Scholarship – Introduction, *Profession* 2011, pp. 123-201, 10.1632/prof.2011.2011.1.123
- Warwick, C; Terras, M. et. al.**(2007) Evaluating Digital Humanities Resources: The LAIRAH Project Checklist and the Internet Shakespeare Editions Project. In: (Proceedings) ELPUB 2007.
- Unsworth, J.** (2001) University of Virginia, Evaluating Digital Scholarship, Promotion & Tenure Cases. http://artsandsciences.virginia.edu/dean/facultyemployment/evaluating_digital_scholarship.html
- UNAM**, Disposiciones Generales para la Actividad Editorial de la UNAM, cited in López, C. and Estrada, A., Edición y Derechos de Autor en las Publicaciones de la UNAM, 2007.
- UNAM**, Recursos web – Lineamientos CATIC. <http://recursosweb.unam.mx/recursos-web/lineamientos-unam/>

LARHRA, France

1. The SyMoGIH project and linked data

The SyMoGIH project has created an open modular platform for storing geo-historical information. The benefit of the platform is to allow researchers to share their data and texts in a collaborative environment. Up to now, about fifty students and scholars are contributing or contributed to the information collection, and ten research projects are using the platform to store data concerning different domains, such as intellectual, economic, social, institutional or religious history. The richness and heterogeneity of the shared information requires a generic data model which was designed with Merise (ERD) modeling method (cf. Beretta, Vernus 2012) and implemented in a relational PostgreSQL database, to which users can connect via a user friendly AJAX web application (cf. Beretta, Vernus & Hours 2012). In parallel, we have recently proposed an environment using eXist-db to share XML/TEI encoded texts. The semantic annotation of named entities and knowledge units (i.e. informations) that we find in the texts is achieved by linking the semantic tags to the resources defined in the relational database.

In carrying out the latter, we realized the importance of identifying the database objects using URLs. In addition to the websites we offer for publishing the different datasets related to specific projects (for instance www.patronsdefrance.fr), we also created a generic website to deliver to the public the whole of our authority files and the knowledge units which the participating scholars decided to make public (www.symogih.org). As a first step, this website provides dereferencing of our URIs in form of a web page with the description of the resource (e.g. www.symogih.org/resource/Actr195 for Johannes Kepler) but it raises the issue of providing dereferencing in form of RDF data and, more widely, of connecting our data to those produced elsewhere. For this purpose, we made an ontology that suits our needs.

2. From a project-specific ontology to a generic semantic model for historians

The generic nature of the SyMoGIH data model allows to easily transpose it to an OWL DL ontology. This project-specific ontology is designed to publish our data and it is based on four main classes : *Object*, *KnowledgeUnit*, *Role* and *Sourcing* (cf. Image 1). An *Object* can be a person, a place, a concept, a bibliographical object, etc. and can be linked to a *KnowledgeUnit* through a *Role*. A superclass *AssociatedObject* gathers *Objects* and *KnowledgeUnits*: this allows a *KnowledgeUnit* to be associated to another one as if it was an object. *KnowledgeUnits* are divided in two subclasses: an *Information* expresses knowledge as it is constructed by the historian using critical method and extracting it from different sources; a *Content* reproduces knowledge as it was meant by one and only one source, even if we know the source is wrong about an event date or circumstances. In both cases, *Sourcing* provides the origin of each knowledge unit. Specific *KnowledgeUnits*' and *Roles*' types appear as instances of the *KnowledgeUnitType* and *RoleType* classes: this means that the ontology is versatile, can be gradually expanded and virtually handles any type of knowledge.

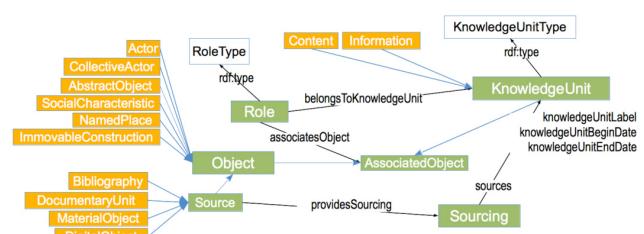


Fig. 1: Classes and properties of the Symogih ontology (An instance like "Johannes Kepler (1571-1594)" belongs to the class <http://symogih.org/ontology/Actor>)

The SyMoGIH project : Sharing and publishing historical and geographical data in a standard, open and interoperable way

Gedzelman, Séverine, Sonia
severine.gedzelman@ens-lyon.fr
Triangle, France

Beretta, Francesco
francesco.beretta@ish-lyon.cnrs.fr
LARHRA, France

Ferhod, Djamel
djamel.ferhod@ish-lyon.cnrs.fr
LARHRA, France

Boschetto, Sylvain
sylvain.boschetto@ish-lyon.cnrs.fr
LARHRA, France

Butez, Claire-Charlotte
charlotte.butez@ish-lyon.cnrs.fr
LARHRA, France

Vernus, Pierre
pierre.vernus@ens-lyon.fr
LARHRA, France

Hours, Bernard
bernard.hours@univ-lyon3.fr

The SyMoGIH ontology bears close resemblance to similar ontologies elaborated by historians and scholars interested in the development of prosopographical databases, particularly the “factoïd” data model developed in King’s College Department of Digital Humanities, London (Bradley/ Pasin, 2013) and the “Aspekte” data model developed by the Personendaten-Repositorium (<http://pdr.bbaw.de/>) (Berlin-Brandenburgische Akademie der Wissenschaften). Although independently conceived, the SyMoGIH ontology shows also close similarities to the Simple Event Model ontology (van Hage et al. 2011) and the TemporalEntity in CIDOC-CRM (Doerr, 2003). This convergence of different semantic models treating time-related information has recently risen a discussion between their designers about the definition of a common ontology for publishing and sharing data produced by historical research.

3. Publishing and querying linked data

Pending the achievement of this very important discussion, we used our ontology to create an RDF view on our collaborative database. The implementation uses D2RQ1 to create a SPARQL endpoint by means of a term map realized according to the RDB2RDF principles2 that rewrites the database structure according to the SyMoGIH ontology. To increase the SPARQL endpoint performance and allow more sophisticated queries, we periodically dump the available data from D2RQ to an OpenLink Virtuoso triple store3 and we use OntoWiki4 for a visual presentation of the dataset. This project is still in experimental phase: at the moment, we are manually interlinking our resources with the one’s found in DBpedia, IDRef5, etc. and we are testing federated queries to visualize and compare informations coming from different datasets. We are also testing semi-automatic interlinking with other data providers, like DBpedia or GND6. This will allow us not only to publish our data on the semantic web but also to compare the quality of available datasets and enrich them, offering new interesting perspectives to researchers in history (cf. Meroño-Peña et al 2013).

References

- Beretta, F., & Vernus, P.** (2012). *Le projet SyMoGIH et la modélisation de l’information: une opération scientifique au service de l’histoire*. Les Carnets du LARHRA, (1), 81–107. (<http://halshs.archives-ouvertes.fr/halshs-00677658>)
- Beretta, F., Vernus, P., & Hours, B.** (2012). *Le Système modulaire de gestion de l’information historique (SyMoGIH): une plateforme collaborative et cumulative de stockage et d’exploitation de l’information géo-historique*. Presented at the Digital Humanities 2012 in Hamburg (http://larhra.ish-lyon.cnrs.fr/iDocuments/SyMoGIH/Poster_SYMOGIH3.jpg.pdf).
- Michele Pasin and John Bradley** (2013). *Factoid-based prosopography and computer ontologies: Towards an integrated approach*, Literary and Linguistic Computing Advance Access published June 29, 2013.
- Doerr, M.** (2003). *The CIDOC CRM - An ontological approach to semantic interoperability of metadata*. AI Magazine, 24, 2003.
- Meroño-Peña, A., van Erp, M., Breure, L., Scharnhorst, A., Schlobach, S., & van Harmelen, F.** (2013). *Semantic Technologies for Historical Research: A Survey*. Semantic Web journal.
- van Hage, W. R., Malaisé, V., Segers, R., Hollink, L., & Schreiber, G.** (2011). *Design and use of the Simple Event Model (SEM)*. Web Semantics: Science, Services and Agents on the World Wide Web, 9(2), 128–136. doi:10.1016/j.websem.2011.03.003
- Vernus, P.** (2009). ANR Système d’information: “*Patrons et patronat français - XI^e - XX^e siècles*” (SIPPAF). Retrieved from <http://halshs.archives-ouvertes.fr/halshs-00474109>
- <http://virtuoso.openlinksw.com/>
- <http://www.idref.fr/autorites/autorites.html>
- <http://www.w3.org/2001/sw/rdf2rml/>

<http://d2rq.org/>
<http://pdr.bbaw.de/>
<http://www.patronsdefrance.fr>
<http://www.loa.istc.cnr.it/DOLCE.html>
<http://symogih.org/ontology/Actor>
<http://www.dnb.de/EN/gnd>

User-friendly lemmatization and morphological annotation of Early New High German manuscripts

Gießler, André

andre.giessler@informatik.uni-halle.de
 Martin-Luther-University Halle-Wittenberg, Institute of Computer Science

Ritter, Jörg

joerg.ritter@informatik.uni-halle.de
 Martin-Luther-University Halle-Wittenberg, Institute of Computer Science

Molitor, Paul

paul.molitor@informatik.uni-halle.de
 Martin-Luther-University Halle-Wittenberg, Institute of Computer Science

Andert, Martin

martin.andert@informatik.uni-halle.de
 Martin-Luther-University Halle-Wittenberg, Institute of Computer Science

Kösser, Sylwia

sylwia.koesser@germanistik.uni-halle.de
 Martin-Luther-University Halle-Wittenberg, Institute of German Studies

Leipold, Aletta

aletta.leipold@germanistik.uni-halle.de
 Martin-Luther-University Halle-Wittenberg, Institute of German Studies

In the humanities the study of ancient texts with multiple conveyed records is confronted with the problem to explicate the relationship among the records and to identify their commonalities and differences. The interdisciplinary project SaDA aims at that task. SaDA stands for "Semiautomatic difference analysis of complex text variants" ¹. It is a BMBF funded project in which germanists, romanists and computer scientists work together. One focus of the project lies on the manuscript "Wundarznei" by Heinrich von Pfalzpaint of the 15th century, of which eleven records are known. The goal is to compare the variants and to display them in a critical edition and an on-line edition with synopsis and critical apparatus. Subject of this work is the philological preparation of the variants.

The way from variants to a critical edition starts with the transcribed manuscripts in Early New High German (ENHG). For a comparison the identification of corresponding words in the variants is crucial. But the German language in this stage lacks a common orthography, which leads to very different spellings of the same word in different variants. Actually, every newly discovered record reveals further spellings. Consequently, a prerequisite for text comparison is the mapping of word forms to signatures, by which corresponding word forms can be found. To get suitable signatures, the word forms are annotated with lemmata and additionally with part-of-speech tags and morphological data such as grammatical case, genus and numeral. The morphological data enables us to map one text more precisely to another text, if they are variants. Using precise query methods, e.g. for the state of certain phoneme building, other texts of this language stage could be dated, localised, dialectal determined or fitted into German language history. Such richly annotated texts are valuable witnesses for studies of ENHG and are therefore planned to be used for open research questions. To this end, accuracy of the annotation data has high priority.

The annotation process is very tiresome, repeatedly lemmata have to be looked up in lexica and grammatical attributes have to be determined. The manual effort is quite high. In the project SaDA software solutions are developed which support humanists by supporting as many procedures as

possible and finally reducing the effort drastically. In this work we present the tool Lemmano, which helps to annotate ENHG texts quickly and intuitively.

From a computer scientists perspective several challenges arise in such a tool. At first the transcriptions have to be evaluated. They consist of UTF-8 encoded text files written by the transcriber. A special notation is used to mark details of the manuscript, such as information about peculiar spellings, diacritics, punctuation, transcribers comments, word separations or relationships between fragments of composed words. The notation follows rules, which have been developed by germanists from Bonn, Bochum and Halle for the encoding of old German texts in a readable and concise manner. A software for evaluating such transcriptions should be able to check the conformance to the transcription rules and to mark errors precisely. Primarily it has to identify the words of the text while recognising noted details such as diacritics, word separations and fragments of composed words. Once recognised a second challenge arises with the annotation of the words. For this task several automatic approaches exist. One approach uses grammatical rules, by which natural language processing tools recognise sentences, segments and phrases, which enables them to conclude part-of-speech tags and lemmata for the words. A second approach is based on lexica by looking up every word form and choosing the listed lemma and morphological data. A third approach is based on probabilities. Here, an annotation tool is trained using annotated exemplary texts. The tool learns word forms with their annotations depending on neighboured words. Once trained it can determine the lemma and annotation for known words in unknown texts with a certain probability.

All these approaches have a major disadvantage: their results are defective because of the complexity and variability of the human language. A further drawback related to ancient manuscripts is their little tolerance for deviating spellings or spelling errors in words, which leads to a drastic decrease of their hit rate. These disadvantages combined with our high demands and unsteady orthography in ENHG manuscripts inhibit their use for the "Wundarznei".

Lemmano's answer is therefore a semiautomatic approach: it presents high-quality suggestions for the current word form to the user by looking for similar word forms in lexica and taking their associated annotations, still allowing the manual annotation of every single word. In contrast to the automatic approaches the choice for the correct lemma and morphological data is made by the user. The user interface is intuitively usable and is designed for massive throughput.

After the import and the check for conformance to the transcription rules the transcribed texts are presented easily readable and without transcription notation in their original line structure. Diacritics and encoded special characters are displayed using appropriate unicode characters. For every word, the user can open a dialogue by a mouse click or pressing the enter key where she can insert, modify and save annotation data. She can either enter all data from scratch or accept or modify a suggestion. The suggestions are computed with the help of lexica available to the tool. For a word form to be annotated the tool searches for "similar" word forms in the lexica, uses their associated annotations and presents them to the user after grouping and sorting.

Let us go into some details using the example of Heinrich von Pfalzpaint's "Wundarznei". There are eleven records known to exist, of which ten are available. The spelling of words in these variants varies that much, such that it is hard to uniquely map them to word forms in lexica. This problem has been addressed previously, for example by² and³ resulting in tools used in⁴. These approaches derive weighted replacement rules learnt by given training data consisting of pairs of original and normalised word forms. Using these rules a list of proposed word forms can be generated for a word form. Lemmano handles this problem similarly, but uses static instead of learnt replacement rules. Lemmano replaces letter combinations in a word by known equivalent letter combinations. For a given word form, a list of variants is created by replacing letter combinations by known equivalent letter combinations. For example, the word form "pfeyl" leads to the variants "pfeil", "pfeil", "pfeijl", "pfejl", "pfeyl", "pffeyl". This approach generally

leads to hits in a lexicon which we have extracted out of the "Bonner Korpus"⁵, an annotated ENHG corpus. To further increase the quality of results Lemmano learns all annotations entered by the user and stores them in a separate lexicon. After annotating a certain amount of text, in more than 60% of the cases the annotation list which is suggested to the user contains the right entry in the first position and in more than 80% in the first or second position.

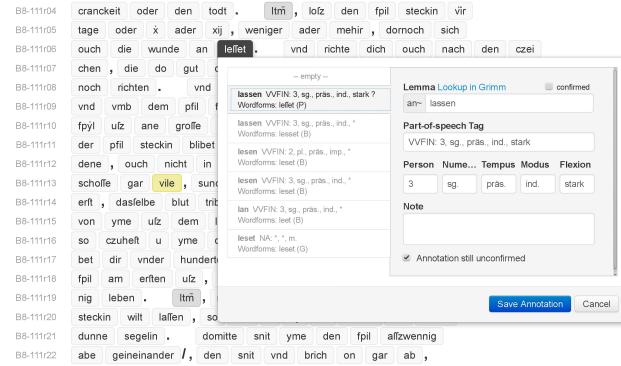


Fig. 1: Annotation dialogue for a word form.

Figure 1 shows the annotation dialogue for a word form which is a fragment of a composed word. On the left in the dialogue box there is the list of suggested annotations consisting of lemma, part-of-speech tag and morphological data.

The manual annotation process can be done quickly, because the user can enter all data with the keyboard. For example selecting words in the text, opening the annotation dialogue and selecting from the suggestion list can be done without losing time with switching between keyboard and mouse. The input fields for lemma, part-of-speech tag and morphological attributes are equipped with auto completion functions, which propose more suitable entries with every entered character. In many cases a word form can be annotated with few keystrokes. The enter key on a word form opens the dialogue, the arrow keys select an entry from the suggestion list, the tab key switches to the input fields to modify or enter new data and a further hit on enter saves the annotation.

In the process of annotating the manuscripts of the "Wundarznei" Lemmano proved itself as a big help. Being a web based tool multiple users can annotate simultaneously and benefit from the learnt annotations of all users.

In the ongoing project, Lemmano is being enriched with further functionality for marking commonalities and differences between variants and with functionality for displaying them as synopsis or critical apparatus. For the identification of corresponding text passages in the variants it uses the annotation data. To this end flexibility of Lemmano has to be extended on the sentence structure, such that text passages with equal content can be found at different places in the manuscripts, even with strong deviations in orthography and grammar.

References

- Project SaDA, Martin-Luther-University Halle-Wittenberg.** (2013). www.informatik.uni-halle.de/ti/forschung/e-humanities/sada (accessed 28 October 2013).
- Reynaert, Martin.** (2011). *Character confusion versus focus word-based correction of spelling and OCR variants in corpora*. International Journal on Document Analysis and Recognition (IJDAR) June 2011, Volume 14, Issue 2, pp 173-187.
- Hauser, Andreas W., Schulz, Klaus U.** (2007). *Unsupervised Learning of Edit Distance Weights for Retrieving Historical Spelling Variations*. Finite State Techniques and Approximate Search. Borovets, Bulgaria.
- EU IMPACT. Improving Access to Text.** (2008-2011). www.impact-project.eu (accessed 28 October 2013). Project funded by European Commission.

5. Diel, Marcel and Fissen, Bernhard and Lenders, Winfried and Schmitz, Hans-Christian. (2002). XML-Kodierung des Bonner Frühneuhochdeutschkorpus, Universität Bonn. www.korpora.org/Fnhd (accessed 28 October 2013).

The MiCLUES system: Dynamic, rich contextual support for museum visits

Gold, Nicolas E.

University College London, United Kingdom

Rossi Rognoni, Gabriele

Royal College of Music, United Kingdom

Introduction

Attracting visitors and maintaining their interest can be difficult to achieve for smaller, specialist museums. Constraints on physical space can create high-density displays which, although good for exposing collections, can make coherent interpretation difficult. Audio or written guides can help but substantial effort is required to generate and maintain them. Their effectiveness is simultaneously hindered by the reduced capacity of concentration that visitors have while standing in a gallery, so the amount of information that can be transmitted through written labels, panels, and audio guides is very limited (Museums Association, 2013). Audio guides are also fairly rigid in the structures and routes they impose owing to the time needed for audio recording and the need to match this with the collection's physical organization. In addition, they allow a solely linear approach to each object (a single story briefly illustrating the item).

Recent work has developed more flexible digital approaches. For example, the Musée de la Musique, Paris (Cité de la Musique, 2013) and the Grassi Museum für Musikinstrumente der Universität Leipzig (Universität Leipzig, 2013) have developed online access to their collection records, and the Museum of Fine Arts, Boston has a digital guide incorporating video, audio, and commentary with the ability to save "favourites" from an online catalogue (Museum of Fine Arts Boston, 2013). The QRator project (Gray et al., 2012) used QR codes to link exhibits to online discussions. Our work shares similar goals to QRator in enhancing the visitor experience. It differs in that our focus is on creating contextualised trajectories through the gallery rather than conversational engagement around the exhibits.

The Museum of the Royal College of Music (RCM) (see Figure 1) was established in 1894 and holds over 1,000 musical instruments, including the earliest surviving keyboard stringed instrument, the earliest known baryton, and the earliest known guitar. It is regularly open to the public, and is visited every year by ca. 5,000 visitors including students and teachers. It is moving to digitise its collection to make it more widely available. This creates exciting opportunities to improve the experience of visitors to the museum itself, through bespoke applications running on smart devices (e.g. Android phones). The RCM galleries currently lack any kind of audio guide or support to the visitor, apart from textual printed labels.



Fig. 1: The RCM Museum (Photo: Chris Christodoulou © RCM)

This poster presents work in progress to provide rich technologically-supported experiences for Museum visitors, offering more meaningful and informative access to the collection and encouragement to further visits. It will also allow the physical space to be used efficiently without harming the interpretive value of the objects on display. The aim is to produce a method that is theoretically and experientially grounded and that can be applied in a wide range of contexts.

Theoretical Approach

Our solution is underpinned by Benford et al.'s conceptual framework of mixed-media performance trajectories (Benford et al., 2009) to design the pathways through the collection. Among other things, the framework allows the formalization of roles, transitions, traversal between physical and virtual experiences, and an episodic structure in the formation of canonical (author-defined), participant (the actual route), and historic (reflective) trajectories (Benford et al. 2009; Fosh et al. 2013). Fosh et al. (2013) present the first proactive (rather than analytical) use of the framework, using it for sculpture garden design. In that instance there was a single canonical trajectory through the experience and low object density. We intend to support a number of canonical pathways (that we term "curated") through higher density displays. Using this framework will enable us to soundly and explicitly address problems found previously in interactive exhibits e.g. multiple interactive features overwhelming visitors (Allen & Gutwill, 2004). Having a clearly defined trajectory will establish an appropriate prioritization and visiting sequence so that even densely populated galleries can be interpreted clearly.

Essentially the pathways offer a guided navigation through the "mesh" of museum resources, artefacts and information, grounded in the physical space of the museum itself, with the physical exhibits as landmarks on the journey. Curated pathways appropriate to the RCM Museum might include musical curiosities, early music, the art of musical instruments (decoration), chronologies, occasions, military music, and unusual materials. By joining parts of the physical and virtual collections into a single tour, visitors benefit from being able to access parts of the collection not physically presented at the time of the visit, and museum staff can benefit from monitoring the demand for such exhibits through their occurrence in popular pathways, bringing them out for physical display in response. The museum's display strategy can thus be informed by user activity at a fine-grained level.

Pathways may be curated by museum staff (canonical trajectories), planned by visitors in advance through a web interface (also canonical but defined by the user) and downloaded to a device, crowd-sourced (visitors could publish pathways for others to download; a kind of historic trajectory) or produced by context-aware recommender-systems. Museum activity can be fed back to visitors through the projection systems currently used to display a slideshow. Heatmap visualisations and pathways followed over a recent period of time could be displayed instead, offering a live "trajectory view" of the collection.

Visitors may, of course, deviate from their planned pathway, e.g. in response to noticing a nearby interesting exhibit, and a guiding system should be capable of adapting to this (i.e. designing for human nature rather than in spite of it (Adams et al., 2004)). Our solution will offer options for returning to the original pathway, not just directly, but through small, thematically-coherent routes (e.g. historical, geographical, stylistic) to provide a richer, user-driven experience of the collection. Semantic metadata will be needed to do this and new algorithms will be developed to search this metadata and deliver the new pathway in real time. One key technical problem is thus the automated, dynamic, real-time design of "micro-trajectories".

Evaluation of the technology will involve museum visitors trying the prototypes and being surveyed on their experiences.

Technological Solutions

We are designing a smart-device app called MiCLUES (Musical instrument Collection articULation for User-driven Exploration with Smart-devices) with two key features:

1. To guide visitors through thematic pathways in the collection, allowing serendipitous diversion and thematically-oriented return to the original path.
2. To give access to digitised forms of the collection elements (including those in storage), extending the context in which the physical and digital artefacts are presented through recorded performances (both historical and modern), documents, animations, and images to provide a rich, dynamic, portable context for a visit.

Connections between artefacts (digital or otherwise) can be drawn at the time of the visit, allowing for the collection to be expanded without requiring costly re-provisioning of the guide itself. Text to speech systems may also alleviate the need to pre-record some or all of the text. Since the software can be made widely available through app stores, the volume (and thus cost) of devices to be maintained by the museum is reduced as visitors can use their own.

To determine where visitors are on their pathways, location tracking will be needed: initially, we plan to use QR codes by each artifact to provide landmarks. These are non-invasive and have been shown to work well in the museum context by the QRator project (Gray et al., 2012) and recent developments in the musical instrument department of the Galleria dell'Accademia in Florence. In the latter case, the codes refer to pages in the online database of text, images, and sound files connected with the object concerned. Although quick to implement and inexpensive, QR codes are poor for accessibility since, for example, visually impaired visitors may not be able to see or scan them and thus may be prevented from using the app. Alternative approaches (e.g. RFID tags) more amenable to tactile interfaces will be explored after the initial prototype is complete to create more accessible ways of following a pathway through the collection. Content in the databases will be encoded to be amenable to screen readers and other assistive technology following appropriate guidance e.g. (Royal National Institute of Blind People, 2005). Other methods of visitor location tracking (e.g. human observation (Guy et al., 2010)) are less appropriate for reasons of practicality, complexity and cost.

Conclusion

This work is developing rich, interactive guides for a specialist museum collection, adopting recent theoretical advances in human-computer interaction to design appropriate trajectories and support these through software capable of redesigning them on the fly. The poster will present the underlying theoretical considerations, practical issues, the app design, and progress gained with early prototypes.

Acknowledgements

This work is primarily supported by Share Academy (an Arts Council England funded programme) and partially supported by the Engineering and Physical Sciences Research Council [grant number EP/G060525/2]. At this stage no data has been produced.

References

- Adams, M., Luke, J. & Moussouri, T.** (2004). *Interactivity: Moving beyond terminology*. Curator: The Museum Journal, pp. 155–170.
- Allen, S. & Gutwill, J.** (2004). *Designing with multiple interactives: Five common pitfalls*. Curator: The Museum Journal, pp. 199–212.
- Benford, S., Ginnachi, G., Koleva, B., Rodden., T.** (2009). *From Interaction to Trajectories: Designing Coherent Journeys Through User Experiences*. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2009). Boston, Massachusetts.
- Cité de la Musique, Paris.** (2013). *Mediatheque: Collections du Musée*. Available at: http://mediatheque.cite-musique.fr/masc/?url=/clientbooklineCIMU/toolkit/p_requests/default-collection-musee.htm [Accessed October 25, 2013].
- Fosh, L., Benford, S. & Reeves, S.** (2013). *See Me, Feel Me, Touch Me, Hear Me: Trajectories and Interpretation in a Sculpture Garden*. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2013). pp. 149–158.
- Gray, S., Ross, C., Hudson-Smith, A., Terras., M., Warwick, C.** (2012). *Enhancing Museum Narratives with the QRator Project: a Tasmanian devil, a Platypus and a Dead Man in a Box*. Proceedings of Museums and the Web 2012. San Diego, CA. Available at: http://www.museumsandtheweb.com/mw2012/papers/enhancing_museum_narratives_with_the_qrator_pr.html
- Guy, G., Dunn, S. & Gold, N.** (2010). *Capturing Visitor Experiences for Study and Preservation*. Proceedings of Digital Humanities 2010. London, UK. pp. 160–163. Available at: <http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/html/ab-716.html>
- Museums Association** (2013). *Exhibition Labels*. Available at: <http://www.museumsassociation.org/museum-practice/exhibition-labels> [Accessed October 25, 2013].
- Museum of Fine Arts Boston** (2013). *MFA Guide*. Available at: <http://www.mfa.org/visit/mfa-guide> [Accessed October 25, 2013].
- Royal National Institute of Blind People** (2005). RNIB guidelines for producers and commissioners of audio guides: making PDA guides accessible for blind and partially sighted people. Available at: www.rnib.org.uk/professionals/Documents/guidelines_audio_guides.doc.
- Universität Leipzig** (2013). *The Museum of Musical Instruments*. Available at: <http://mfm.uni-leipzig.de/en/dasmuseum/Geschichte.php> [Accessed October 25, 2013].

How a story is performed: traditional storytelling in the hands of computing

Gomes, Mariana

mariana.gomes@clul.ul.pt

Linguistics Centre of the University of Lisbon

General description

The main idea is to find out how to edit a collection of oral transmitted texts of literary interest (often called orature), in the future resulting as a marked-up database, serving as a model to future researchers and projects. The main goal is to analyse both the transcribed texts and the way they are performed

orally. Until today, several research areas focus on particular parts of this kind of enunciation – Literary studies concentrate on the transcription, Anthropologic studies concentrate on the artistic performance or on the interactants' behavior, etc. But nowadays we have the means that allow us to study a more complete contextualization of an oral composition, and interaction including voice characteristics (intonation, prosody), body movements (face, torso, arms, hands) and proxemic behaviour (the relation between an interviewee and the interviewer). My presentation will focus on the main parts of my PhD project: a) the creation of a representative corpus of Portuguese oral tradition with performative and literary annotation, b) digital publication of a more complete edition model, and c) gesture and literary classification.

Previous Studies

From previous incipient work done on oral prayers themed on the Passion of the Christ, it proved to be productive to assume that performance integrates verbal art in its many strands. Performative elements have given some hints on their importance, having this previous study focus on the intonational contrast between the informant's enunciation of a prayer and in spontaneous speech.

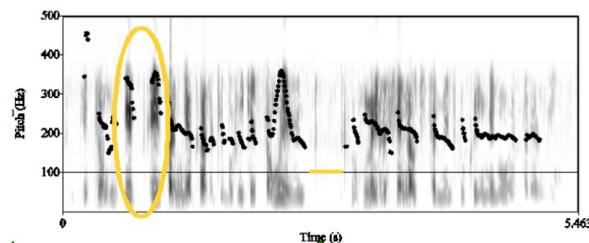


Fig. 1: Image A – spontaneous speech

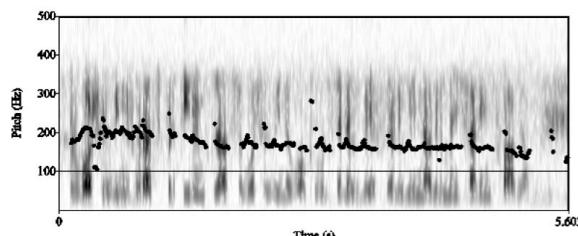


Fig. 2: Image B – enunciation of prayer a)

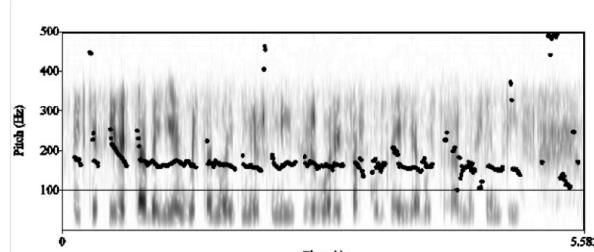


Fig. 3: Image C – enunciation of prayer b)

In image A ascending and descending melodic variations contours can be seen, characteristic of human speech. In images B and C, F0, paradoxically, is almost flat, indicating a type of monotonous speech, showing a lowering of the medium tone of voice of the speaker in question. Moreover, the total duration of each one of the extracts, the informant produces a greater amount of words during prayer enunciation than in conversation mode in which, as indicated in A, disfluencies or pauses occur, which rarely occur during prayer enunciation. These characteristics are due to the fact that, among other reasons, discourse planning is slower during conversation than

during prayer enunciation (being a discourse cited by memory, not involving simultaneous processing to the production).

Concluding, the above-explained study shows these to be performative marks to the production of a prayer orally and it gives us hints to question whether other types of orally transmitted texts have them or not.

Data collection and tools used

The data is being collected through fieldwork interviews and archival research, which result in the primary database in video and audio formats that share a common structure for storage. Each entry is accompanied by metadata related to the context of collection: informant's data, place and date, technical information about support files. The data is worked to ensure time alignment of the different types of representation: transcription of the text, kinesic and verbal (not literary) markup. To this end, the programs being used until now are 'EXMARaLDA' – for the alignment of text and video annotation –, and 'Praat' for intonation and prosodic analysis of the compositions. Together, these elements constitute the support for the database analysis: literary classification according to the taxonomic reference indexes, annotation of text versions and variants, motifs (using AMICUS network labelling), UNESCO intangible heritage taxonomy and Matriz PCI, formal characterization of the text (composition, data contamination or counterfeiting, domain – religious, artistic, etc. – fictionality degree), linked existing Wikipedia entries for each type of text, and literary and compositional context. This way, the database will gather several interpretations on the same composition, letting the user/researcher work with the information that suits their interests.

The Early Modern OCR Project (eMOP): Fostering Access to Early Modern Cultural Materials

Grumbach, Elizabeth

egrumbac@tamu.edu

Texas A&M University

Mandell, Laura

mandell@tamu.edu

Texas A&M University

Christy, Matthew

mchristy@tamu.edu

Texas A&M University

1. Introduction

The Initiative for Digital Humanities, Media, and Culture at Texas A&M University received a \$734,000 grant from the Andrew W. Mellon Foundation in 2012 to make machine readable 45 million pages of data¹. By partnering with Gale and Proquest, eMOP will combine open source OCR (Optical Character Recognition) software and book history in order to improve the accuracy of

OCR for early modern (1473-1800) texts². The Early Modern OCR Project (eMOP) aims to publish an open source OCR workflow, improve the visibility of early modern texts by making them fully searchable³, and form a community of scholars and institutions interested in the digital preservation of these texts⁴. Our goal is to foster collaboration among various disciplines, and, in doing so, cultivate inter-institutional and international relationships that make possible new kinds of humanities research.

2. Poster Description

Our workflow (see images below for two early drafts of our workflow, subject to change) blends the disciplines of book history, digital humanities, textual analysis, and machine learning in order to create a corpus of keyed texts that are far more correct than is now possible with the current set of tools. These keyed texts will improve access to early modern texts that are currently only searchable through “dirty” OCR or metadata alone. The open source OCR workflow will contain, among other things, access to an early modern font database, customization guidelines for the Tesseract OCR engine, post-processing and diagnostic algorithms, and crowdsourcing correction tools.

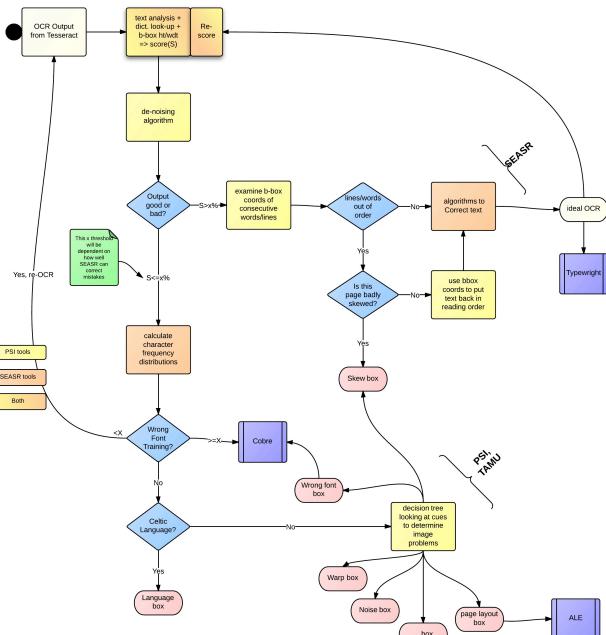


Fig. 1: Two versions of the eMOP workflow from October 2012 and February 2014, respectively.

In addition to presenting a detailed and accurate representation of our OCR workflow for early modern texts, we intend to present the following aspects of eMOP:

- Information on how to obtain the open source code for all of the tools, software, and workflows that eMOP has produced.
- How our tools and software can be used by individual scholars, instructors, and institutions in the classroom, for an OCR project, or for personal research.

3. Demonstration Description

We intend to go beyond presenting an overview of the project; instead our poster and demonstration will communicate the concrete solutions found and software available to address the “OCR problem” from a Digital Humanities perspective. To this end, we will demo our five crowdsourcing and scholar-sourcing correction tools for conference attendees. These demonstrations of the tools will be operating in production with our eMOP OCR output of the EEBO ECCO datasets (45 million pages).

- The Franken+ tool, developed by Bryan Tarpley at Texas A&M, enables the creation of an “ideal” typeface using glyphs identified in scanned images of documents from the early modern period. Franken+ also exports these typefaces to a training library for the open-source OCR engine Tesseract⁵.
- Aletheia Layout Editor (ALE), developed by PRImA at the University of Salford, is a crowd-sourced correction tool for re-drawing regions on problematic OCR’d pages, such as Title pages, multi-columned texts, image-heavy documents, and more.
- The TypeWright software, developed by Performant Software Solutions and 18thConnect, enables users to

correct the “dirty” OCR of an entire early modern document, and our partnership with ECCO allows 18thConnect to release fully corrected documents to their scholar-editors in plain text and TEI-A formats.

- The Cobre tool, developed by Dr. Anton DuPlessis and Cushing Memorial Library at Texas A&M, enables scholar-experts to compare, re-order pages, and annotate the metadata for multiple printings of documents in the eMOP dataset.
- The Anachronaut tool, developed by a team of undergraduates and Dr. Ricardo Gutierrez-Osuna at Texas A&M, is a Facebook game that uses the power of Facebook (and many layers of user confidence testing) to correct single words and phrases.

References

1. **Mandell, Laura.** *Mellon Foundation Grant Proposal: “OCR’ing Early Modern Texts.”* Grant Proposal. 30 Jun 2012.
2. **Heil, Jacob and Todd Samuelson.** *Book History in the Early Modern OCR Project, or, Bringing Balance to the Force.* Journal for Early Modern Cultural Studies 13.4 (2013): 90-103. Web. 30 Oct 2013.
3. **Mandell, Laura.** *Digitizing the Archive: The Necessity of an ‘Early Modern’ Period.* Journal for Early Modern Cultural Studies 13.2 (2013): 83-92.
4. **European Commission: The Comité des Sages. The New Renaissance: Report of the comité des sages on bringing Eu rope’s cultural heritage online.** By Elizabeth Niggemann, et al. 10 Jan 2011.
5. **Katayoun, Torabi, Jessica Durgan and Bryan Tarpley.** *Early modern OCR project (eMOP) at Texas A&M University: using Aletheia to train Tesseract.* Proceedings of the 2013 ACM symposium on Document Engineering. New York: ACM, 2013.

The Annotated Star: A Collaborative Digital Edition of Rosenzweig’s Star of Redemption

Handelman, Matthew

handelm@msu.edu

Michigan State University

Wygoda, Ynon

ynon.wygoda@mail.huji.ac.il

Hebrew University, Yale University

Rojansky, Shay

roji@roji.org

Eagle TS R&D

Rusinek, Sinai

sinair@vanleer.org.il

Van Leer Institute Jerusalem

For the Digital Humanities Conference 2014, we propose to present a poster discussing our project, “The Annotated Star”: a collaborative, dynamic edition that combines the traditional critical edition with the potential of digital technologies.¹ The poster will draw on our international collaboration developing an open-source, web-based text annotation interface and explore the practical and theoretical implications of our collaborative editing and computer-aided quote-mining of Franz Rosenzweig’s (1886-1929) philosophical-theological monograph *Star of Redemption* (1921). Drawing on a host of interdisciplinary, interfaith and stylistic sources from the German and Jewish cultural tradition, religious texts, and the natural sciences, Rosenzweig’s writings not only serve as an exceptional model - in form as well as in content - for the future of digital textual collaboration.² Even more so, as we hope to demonstrate, the digital editing of Rosenzweig’s text also serves as the impetus for a discussion over how computer technologies can help us rethink, reposition and enhance the

traditional hermeneutic task of close and critical reading in the digital age.

The poster and our presentation will center on the two key aspects of our project. First, they will outline our construction of an online, collective annotation graphical user interface, designed to allow users to upload their own comments, links, and annotations to the digitized text. At first based on Rosenzweig's *Star of Redemption*, such a website will be easily adaptable for other base texts and will link, display, and organize digitized versions of the reference texts from which the base text quotes. The finished site will store previously marked annotations, allow for user input, and display linked texts, as the following screenshots illustrate:



Fig. 1: Prototype Screenshots of the GUI

Such a platform enables and stores collaboratively defined, hyper-textual linkages between *Star of Redemption* and, for instance, Goethe's *Faust* or Kant's *Critique of Pure Reason* – if not for any text within, or even beyond, the humanities. Yet the parallel display of texts enabled by digital technology does not supplant our traditional understanding of close reading, but rather enhances close reading, while visualizing the processes of textual reference and reassembly that serve at its foundation.

Second, our poster and presentation will discuss our implementation of an automatic text re-use detection tool, which we are currently using to detect textual overlaps between *Star of Redemption* and the growing canon of digitized literature, philosophy and the natural sciences. With this tool we hope to use information technologies to unearth moments in which Rosenzweig quotes from sources without citing them and, thus, to provide automated data for the collaborative website discussed above. As our poster explores, the upshot of the project is the discovery of new references in *Star of Redemption* and, with them, a deeper understanding of the rich layers of philosophical, Rabbinic, literary and cultural commentary hidden yet ubiquitous in Rosenzweig's text. Furthermore, it would allow us to revisit, reformulate and re-map the networks of intertextuality and textual re-use that, at their core, inform our hermeneutic and humanistic questions in the present.

Ultimately, we hope to provide more than just a demonstration of a collective critical edition of Rosenzweig's *Star of Redemption*, more, too, than an open-source annotation and editing tool. Rather, by looking at how developments in digital technologies and the Digital Humanities aid in our understanding of a text key to both German and Jewish Studies, it provides us the opportunity to address questions central to the evolving field of the Digital Humanities. Our poster on the "Annotated Star," in other words, will explore how the digital revolution redefines and simultaneously solidifies the task of humanistic scholarship. Perhaps most importantly, it also explores and seeks to engender discussion around numerous philosophical concepts – such as Rosenzweig's ideas on truth and collaboration, translation and language – that enable us to think critically through the new possibilities the Digital Humanities offer for humanistic scholarship and inquiry.

References

- For the history and typology of dynamic textual editing, see **Ray Siemens et al.**, *Towards Modeling the Social Edition*, in Literary and Linguistic Computing, (2012): 445-61, especially 446-9.
- Peter E. Gordon, Rosenzweig and Heidegger** (Berkeley: University of California Press, 2003), **Mara Benjamin**, Rosenzweig's Bible (Cambridge: Cambridge University Press, 2009), the contributions to the *Rosenzweig Yearbook*, vol 2:

Criticism of Islam, ed. Martin Brasser (Freiburg: Karl Alber, 2007).

The SMART-GS Project: An Approach to Image-based Digital Humanities

Hashimoto, Yuta

yhashimoto1984@gmail.com
Kyoto University

Aihara, Kenro

kenro.aihara@nii.ac.jp
National Institute of Informatics, Japan

Hayashi, Susumu

susumu@shayashi.jp
Kyoto University

Kukita, Minao

minaokukita@gmail.com
Kyoto University

Ohura, Makoto

Makoto.Ohura@mb4.seikyou.ne.jp
Kyoto University

Introduction: Difficulties in Working with Handwritten Manuscripts

Some of historically important manuscripts, especially those written in the modern age, are hard to read due to their authors' unclear handwriting. Transcription processes for these manuscripts tend to be more time-consuming, eventually decreasing historians' productivity. When manuscripts are written in East-Asian languages such as Japanese, which have a vast number of characters, transcriptions are even harder. Because historians of the modern age face the challenge of studying a large number of documents, these difficulties can be crucial to them.

SMART-GS, a system for image-based historical studies, has been developed by Japanese historians and developers since 2006 to help historians work on such manuscripts, and has been successfully applied to six historical research projects¹. It is written in Java, and runs on Windows, OS X, and Linux. Its source code is distributed on sourceforge.jp under the GPL 2.0 license². In this paper, we will discuss the approach the SMART-GS project has taken, and its applications to historical studies of handwritten documents.

Project Background

SMART-GS was originally developed by Susumu Hayashi for his study of the history of 19th century mathematics. It was first built for helping the analysis of the mathematical notebooks of the prominent mathematician David Hilbert. These notebooks consist of short notes, each expressing Hilbert's mathematical ideas. In order to identify the time of writing of the notes, Hayashi developed a system that supports image-and-text markup, one-to-many links between markups, search of handwritten texts, and so on. Later the system, named SMART-GS, turned out to be applicable to other historical studies and has been adopted so far by six research programs for the analysis of handwritten manuscripts.

There are a number of applications and web services which look similar to SMART-GS, such as Image-Markup-Tool³ and T-PEN⁴. However, they are mainly aimed at creating transcriptions or annotations for archival purposes, whereas SMART-GS is designed for a different purpose: to streamline historians' work flows, making them more productive.

Supported Features

One of the most basic features of SMART-GS is its markup system for texts and images. SMART-GS can markup an image region in various ways: selecting it with a rectangle or polygon shape, putting comments on it, drawing an arrow from one region to another, and so on (See Fig. 1). This markup information will be stored in an XML file separately from the original images. Therefore a user only has to exchange a small XML file to share his or her project with others; provided they have the same image files. In addition, it's possible to create a bidirectional and one-to-many link between any two markups. This feature enables users to make correspondences between an image region and its transcribed text.

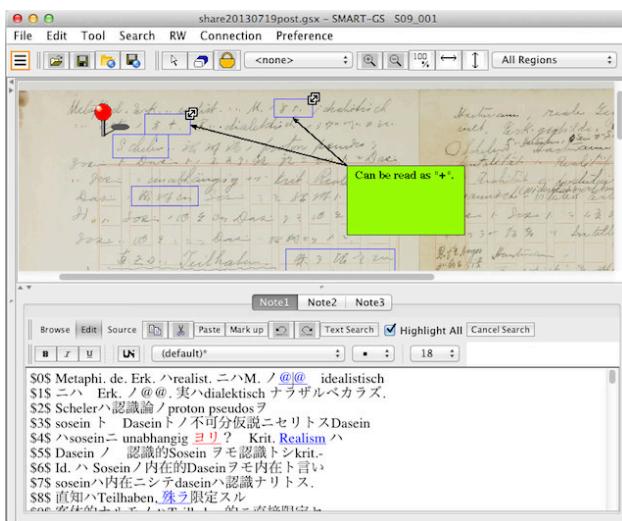


Fig. 1: Markups for image and text

For deciphering illegible words in manuscripts SMART-GS features an image search function. Using this feature users can easily find a word or phrase that has a similar shape to the query image (See Fig. 2). Also, SMART-GS supports adaptive-repetitive search, in which users can recursively increase the relevance of search results.

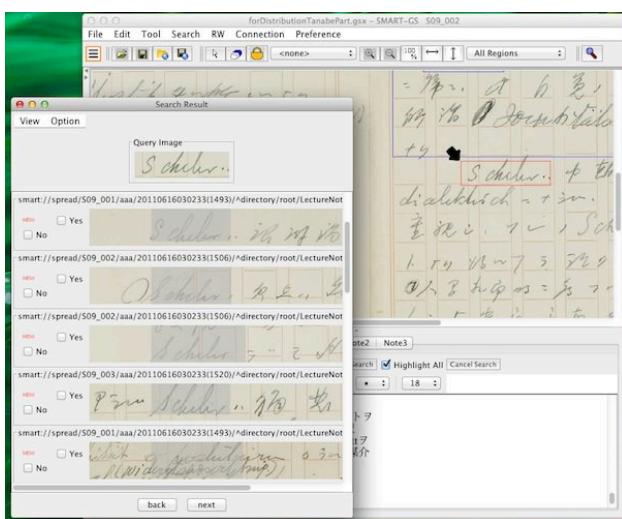


Fig. 2: Search results for an image query "Scheler", a German philosopher's name

For team-based projects SMART-GS offers resource sharing through the Internet. Metadata added on documents and their revision histories can be uploaded to and saved in the subversion-like version control server HCP (Humanities Cyber Platform), so that the project members can work together on the same documents (further cloud collaboration support is planned).

Conclusion

Historical studies based on handwritten texts generally require considerable time and experiences. SMART-GS' features such as image-and-text markup, image search, and project sharing through the Internet, can streamline historians' workflows so that they can focus on the analysis of the content in the manuscripts.

Currently we are implementing two other features into SMART-GS: a built-in TEI editor and automatic mass transcription function. Since at present SMART-GS only supports relatively simple HTML markup for texts, the semantic richness of TEI will bring more power of expression to its markup and link system. The automatic mass transcription will be realized as an application of SMART-GS' image search function. This feature will make it much more efficient to transcribe texts to which OCR is not applicable.

References

1. SMART-GS was for the first time presented in Hayashi, S. (2007). SMART-GS Project: a tool for searching, marking up, and linking historical documents. A talk delivered in the 3rd international symposium of Grant-in-Aid for Scientific Research Areas "Japanese Technological Innovations". sts.kahaku.go.jp/tokutei/intforumprogram_No2.php (Japanese)
2. en.sourceforge.net/projects/smart-gs/
3. tapor.uvic.ca/~mholmes/image_markup/
4. t-pen.org/TPEN/

Open-Access Cultural Heritage Resources and Native American Stakeholders: A Case Study from Chaco Canyon, New Mexico

Heitman, Carrie C.

University of Nebraska-Lincoln, United States of America

Digital Archives are typically designed to address the needs of specific audiences. In the case of cultural heritage resources, the stakeholders invested in digital archival resources can be harder to gauge and further complicated by historical, economic, religious, political and legislative issues. In this poster, I present a case study profiling these complexities: an archive of historic or legacy data on the cultural heritage of a sacred Ancestral Pueblo site known today as "Chaco Canyon" (located in northwest New Mexico, U.S.).

The poster is framed in three sections. Part one documents the ways in which academics and cultural heritage professionals tried and failed to engage native communities when initially building the Chaco Digital Initiative. I highlight some of the reasons those efforts were unsuccessful and the ways in which we succeeded by engaging in the process of tribal consultation. I also discuss some of the reasons our project was ultimately accepted if not embraced by various tribal representatives. This case study is presented in the context of current international efforts, such as Mukurtu, to create community-based repositories for cultural knowledge as well as issues of digital repatriation.

In the second part of the poster, I outline the culturally sensitive issues we encountered during the course of the project and the technical and social solutions we used to address them. Such issues include withholding images of human remains from image searches and decisions about how to handle documents containing drawings of human burials. In this section, I briefly describe our engagement in legal disputes and repatriation claims related to the Native American Grave Protection and Repatriation Act passed by the United States Congress in 1990.

In the third part of the poster, I discuss ongoing efforts to engage with descendant communities in ways that allow the

Chaco Research Archive (CRA) to be a resource for cultural empowerment. Structural inequalities facing tribal historic preservation offices often hinder their ability to access and utilize resources like the CRA. I outline current efforts to engage in a more active dialogue and outreach activities to facilitate the use of this digital archive by tribal representatives.

To conclude, I summarize the lessons we have learned through the process of building a digital archive devoted to Indigenous cultural heritage and provide recommendations for best practices.

Collaborative Scholarly Building with the Early Caribbean Digital Archive

Hopwood, Elizabeth

hopwood.el@husky.neu.edu

Northeastern University

Doyle, Benjamin

doyle.ben@husky.neu.edu

Northeastern University

This poster presentation will introduce scholars to the collaborative components of The Early Caribbean Digital Archive, a project of Northeastern University's NULab for Texts, Maps, and Networks.

The Early Caribbean Digital Archive (ECDA) is a highly interactive digital scholarly lab for the collaborative research and study of pre-20th century Caribbean literature. The ECDA seeks to engage scholars and students in a shared, critical study of the textual, material, and cultural histories of the Caribbean by providing them innovative digital technologies and platforms for generating new and understudied knowledges of the Caribbean's rich body of materials. Our approach to this digital archive solves major challenges facing scholars of Caribbean literature; currently no such pan-Caribbean digital or analogue archive of pre-20th century materials exists. This continued absence of a robust digital archive has largely been the result of the history of empire and colonialism in the Caribbean region, where the negative longstanding impacts of imperialist/colonialist practices is visible in the fragmentation and division of Caribbean print materials among archives in Europe, North America, and the Caribbean. Writings from the Caribbean offer critical perspectives of the broad movements of history and culture in the Atlantic world. In an effort to respond immediately to the needs and imperatives of the wide range of scholars already engaged in generative, collaborative scholarly work in the interdisciplinary field of Caribbean studies, scholars at Northeastern University's NULab for Texts, Maps, and Networks began building the ECDA in the spring of 2012.

Our site will foster a shared and informed engagement with the Caribbean and its literary, aesthetic, cultural, and political impact on the study of the pre-C20th century Atlantic world. The project will not only preserve original texts, but will also reframe the literary history of the early Caribbean as one where something new is preserved—voices beyond the imperial history of the Caribbean.

Our poster presentation will emphasize how the archive functions as a working lab through which a diverse population can not only access materials but interact and use them while developing and building the project as a whole. For instance, the site allows image annotation and transcription functions that allows conversations between scholars about the materials. The site is currently hosted at <http://scholarscommons.neu.edu/omekasites/ECA>. This Omeka installation is the first phase in the ongoing development of the ECDA's digital text analytics research lab. The ECDA project, broadly speaking, seeks to establish strong partnerships with a wide range of publics interested in developing digital analytics, digitization techniques, and digital research methodologies that will benefit a shared and informed engagement with the Caribbean and its literary, aesthetic, cultural, and political impact on the study of the pre-C20th century Atlantic world. Therefore, the creation of

the ECDA will provide scholars, teachers, and students an immediate opportunity to begin working with these valuable Caribbean materials, where they will participate directly in the collaborative building of both the ECDA digital archive project and a shared Caribbean studies discourse and scholarly practice.

The digitization of these materials serves an ethical imperative for making these important cultural histories and otherwise difficult to access materials available to a necessarily broad and critically engaged audience. The practice of digitizing and performing digital analyses of these materials raises important questions about both digital humanities practices and methodologies as well as practical questions regarding the establishing of cross-cultural, transnational, multi-institutional, transdisciplinary partnerships in the building of such a massive project. Supported by NULab for Texts, Maps, and Networks, we have also partnered with the Digital Library of the Caribbean.

Our session will invite collaboration and discussion on how the ECDA can make possible not only the compiling and transcription of primary source materials, but also innovative research possibilities for bibliographic and political histories of the site's catalogued items; provocative and educational developing and designing of curated exhibits; and an open-access and interactive platform for offering original research and analyses of personal, economic, political, and cultural histories, all of which will contribute to Caribbean studies discourse and scholarly practice.

The Development of The Dickens Lexicon Digital and its Practical Use for the Study of Late Modern English

Hori, Masahiro

Kumamoto Gakuen University, Japan

Imahayashi, Osamu

Hiroshima University, Japan

Tabata, Tomoji

Osaka University, Japan

Koguchi, Keisuke

Yasuda Women's University, Japan

Nishio, Miyuki

Kinki University, Japan

Nagasaki, Kiyonori

International Institute for Digital Humanities, Japan

Our project to create *The Dickens Lexicon Digital* is based on the index cards which the late Dr. Tadao Yamamoto (1904-91) compiled. Dr. Yamamoto first conceived a plan for the compilation of the *Dickens Lexicon*, and published *Growth and System of the Language of Dickens: An Introduction to A Dickens Lexicon* in 1950, as an introduction to the *Dickens Lexicon*. In 1953 he was awarded the Japan Academy Prize for this work. In order to compile the *Dickens Lexicon* he collected the materials not only from all of Dickens's works but also from his letters and speeches. Unfortunately, in 1991 he died without seeing his vision accomplished.

The project for the *Dickens Lexicon* was organized in 1998 by a research group of 20 (presently 22) scholars. Our ultimate aim has been to compile the *Dickens Lexicon* from approximately 60,000 cards, which Dr. Yamamoto elaborately drew up and left to us, believing that even if the work was not issued in his lifetime, his pupils and successors would be able to publish it at some future date.

Our *Dickens Lexicon* is neither book-based nor document-based, but is rather designed as a web-based reference

resource. Users will be able to search and retrieve lexical data (an idiom, definition, source, quotation, and notes), stored in the original card-database of approximately 60,000 indexed entries, without the need to install extra software (apart from a web browser) on their computer. Some of the types of information on Dickensian idioms should prove quite valuable for non-native researchers of English in particular, as certain idiomatic expressions in English which are common to native-speakers of English may be not noticed as idioms, or not understood as ironical or collapsed idioms.

Fig. 1 shows a result of the retrieval of idioms beginning with the verb “do” in a test version of *The Dickens Lexicon Digital*. The 199 examples in Fig. 1 are ordered alphabetically but can also be listed in the chronological or alphabetical order of the titles. Moreover, as Fig. 1 illustrates, if you want to know more information about the idiom “do the honours of the house,” clicking on the idiom will return a definition drawn from the *Oxford English Dictionary* or any other dictionary, the text and the context where the idiom is used, comments by Dr. Yamamoto and relevant notes from reference materials referring to the idiom.

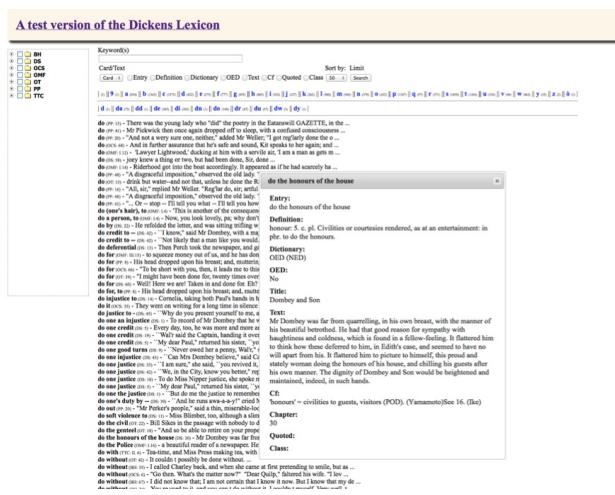


Fig. 1: A test version of The Dickens Lexicon Digital

The lexicon will also be implemented with a multifunctional information retrieval system. In addition to the indexed entries, the lexicon will make it possible to retrieve frequency information on lexical items (from single words to phrases, including multi-word units) drawing upon the full corpus of Dickens' texts, and an additional set of almost comprehensive 18th- and 19th-century fictional texts. A range of functions including concordance display, sort capability, distribution charts, and statistical data with t-score and MI-score and log-likelihood ratios will be available in a user-friendly interface. Therefore, a close scrutiny of idioms appearing in the *Dickens Lexicon*, incorporating this multifunctional information retrieval system, will not only make us more aware of the ways idioms represent an important facet of Dickens' usage of English (compared with those in almost comprehensive 18th- and 19th-century fictional texts), but will also provide greater insight into the characteristic structure of idiomaticity in the English language.

The Dickens Lexicon is expected to be released as *The Dickens Lexicon Digital*, an Internet website with a multifunctional search engine. It will be able to contribute a range of research topics including the following:

- (1)A study of the language and style of Charles Dickens.
 - (2)A comparative study of idiomatic expressions between Dickens and other writers in the 18th and 19th centuries.
 - (3)A chronological research of idioms.
 - research into the language and style of 18th- and 19th-century fictional texts.
 - (5)A research into the history of the English language.
 - (6)An interdisciplinary research between information science and language study.

The Dickens Lexicon Digital is scheduled to be available on the web by March 2017, although it will be partially functional and available on the web in May 2014.

References

- Imahayashi, Osamu, Masahiro Hori, Akiyuki Jimura, and Tomoji Tabata** (2008) "The Dickens Lexicon Project."ERA, Vol.25, The English Research Association of Hiroshima, 43-53.

Yamamoto, Tadao (1950 [2003]). *Growth and System of the Language of Dickens: An Introduction to A Dickens Lexicon*. Hiroshima, Japan: Keisuishsha.

Yamamoto, Tadao (1954) "A Memoir of the Joint Research for the Compilation of the Dickens Lexicon," *Anglica*, Vol. 1, No. 5, 438-51.

What we make of Code: The Role of Programming in the Digital Humanities

Jakacki, Diana

Jakacki, Diane
diane.jakacki@bucknell.edu
Bucknell University

O'Sullivan, James
josullivan.c@gmail.com
University College Cork

1. Introduction

Digital Humanities remains something of an embryonic field; precise definitions of its multifaceted aspects still very much open to debate. The role of software development has proved problematic, with scholars divided on the level with which digital humanists should be actively engaged with programming. Using both quantitative and qualitative methods, this paper seeks to present a diverse range of perspectives from within the Digital Humanities community, all of which address the question: "To be a digital humanist, do you need to be building things?"

By surveying active members of our community, this paper will present findings on development practices within DH scholarship, specifically in relation to what technologies are being used, how they are being deployed, and what geographical and cultural differences, if any, exist. In taking such an approach, this paper will not seek to offer a novel definition of the field of Digital Humanities, but rather, present some objective findings on relevant attitudes in relation to development within DH projects. In doing so, this paper will present the first complete and specific study of the role of software development and programming in the Digital Humanities, developed through the responses of researchers, teachers and practitioners from across the community. Our preliminary findings have indicated that traditional understandings about technical competencies among digital humanists do not bear out, and that commonly held paradigms need revisiting. Presumptions surrounding technical competencies and attitudes, particularly in relation to age and formal qualifications, do not hold true.

This issue was raised by Stephen Ramsey at the 2011 MLA. Ramsey remarked in his paper, "Who's In and Who's Out",¹ that "Digital Humanities is not some airy Lyceum. It is a series of concrete instantiations involving money, students, funding agencies, big schools, little schools, programs, curricula, old guards, new guards, gatekeepers, and prestige. It might be more than these things, but it cannot not be these things." He then puts forward the question, "Do you have to know how to code?" His answer is clear: "I'm a tenured professor of digital humanities and I say 'yes.' So if you come to my program, you're going to have to learn to do that eventually." This paper is will present the results of the first survey to invite members of

the community to give their attitudes on code and its necessity to scholars who describe themselves as digital.

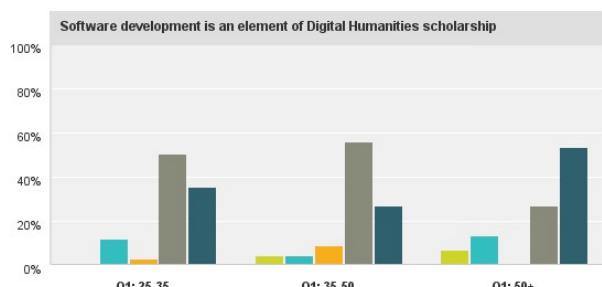
2. Methodology & Initial Findings

Our study avails of a mixed method approach, with our 96 participants, all of which were identified as being actively engaged in the Digital Humanities, responding to a series of quantitative and qualitative questions. The purpose of these questions was to establish, firstly, the level to which Digital Humanities scholars were actively engaged in programming, and secondly, how they view the importance of any such engagement. Questions were divided between two general types: those which asked respondents to give their views on the relevant issues, and those which challenged users to explain their understanding of generic technical details. The purpose of the latter was to help discern if Digital Humanities scholars could demonstrate a fundamental understanding of some of the key terms associated with programming. Before analysing responses as a complete set, we filtered respondents by demographic information that we considered to be of interest, specifically age and gender.

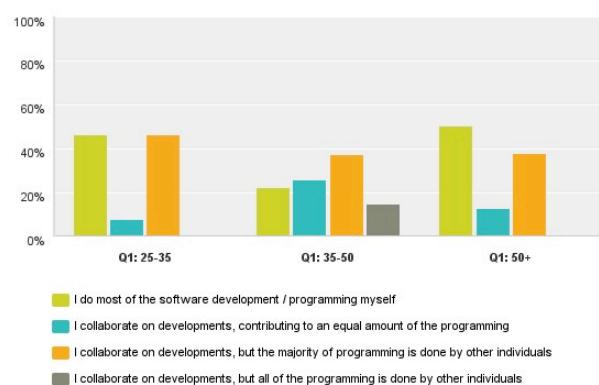
2.1 Age

We were curious to filter respondents by age in an effort to explore any correlations between the practices of scholars and what is often perceived as “generational differences”. The commonly held notion of digital natives suggests that a new generation of scholars is transforming the academy as a result of their increasing familiarity with technology. Our results demonstrate that this is a naïve assumption, with the responses from our 25-35, 35-50 and 50+ age range offering some interesting points of comparison. We only had two respondents in the 18-25 category.

In particular, when asked if “software development is an element of Digital Humanities scholarship”, we found that the majority of respondents over 50 “strongly agreed” that such was the case, as outlined in Figure 1. This contrasts with the other two groups, the majority of which only “agreed” with the aforementioned statement.



When asked to indicate “the type of development practices with which they are most frequently engaged”, as shown in Figure 2, the majority of the 50+ group stated that they do most of the programming themselves, while the majority of the remaining groups either contribute an equal amount to collaborative developments, or have other individuals do the bulk of the project’s coding.



Our findings suggest that the sense that established scholars are more entrenched in traditional views is naïve. The general expectation is that younger scholars, as a result of their perceived familiarity with technology, may be more inclined to take on the development aspects of projects themselves. Our findings demonstrate that the opposite is the case, with the 50+ group being the most self-sufficient in relation to more technical activities. There are some interesting interpretations on academic culture that will be teased out on this point. These findings could arguably be the product of younger scholars having come up in an interdisciplinary environment and having a genuine appetite for collaboration, with the older generation indicating that they prefer a “traditional”, more isolated approach to research. Alternatively, digital natives may not be as technically proficient as many commentators suggest. Technological ubiquity has led to a new generation of scholars who are increasingly familiar with consumer electronics and intuitive graphical interfaces, but our results suggest that these scholars are avoiding more complex technical challenges. This is perhaps supported by a question later in the survey, which finds that of the 25 – 35 age group, the majority of respondents admitted that they did not consider themselves technically proficient. These results might also be representative of a changing academic culture, whereby students are demanding increasing supports from their institutions. In the relevant qualitative portions, it was clear that this age group connected “learning” to “privilege”, in the sense that technical expertise were reserved for scholars with access to appropriate support from their universities. The older groups, conversely, cited the need for scholars to pursue independent development of their technical skills.

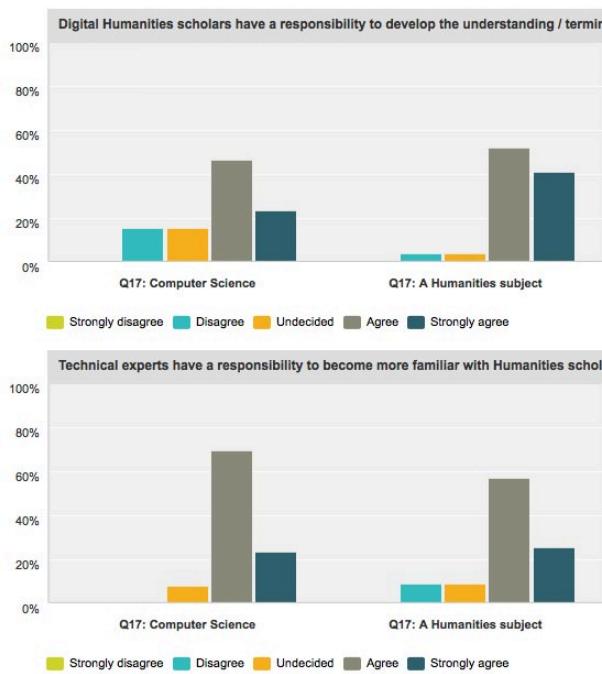
2.2 Gender

There were few distinctions between respondents based on gender, the only significant separation being in relation to the use of software development as an element of one’s work. Considerably more males claim that software development is an aspect of their work, and furthermore, consider themselves technically proficient. This survey, of course, is no indication that this is actually the case, although there were a higher proportion of male respondents possessing formal qualifications in technical subjects. It would be worth comparing these results with data from wider technical disciplines and industries, to see if they are merely a symptom of a wider situation.

2.3 Collaboration

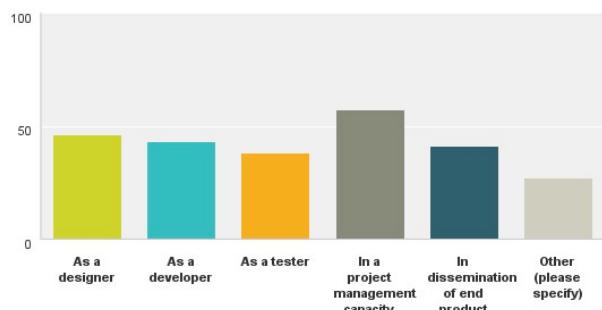
When all respondents were taken as a single set, the key themes to emerge from both the qualitative and quantitative data was collaboration. Most significant in this respect was arguably the way in which respondents with technical expertise expressed a conscious desire to understand the requirements of Humanities scholars, and vice versa. This was somewhat surprising, as the expectation was such that collaborators from across differing fields would have a vested interest in their own areas. As can be seen from Figure 3, respondents possessing formal qualifications in Computer Science agreed that they

have a responsibility to understand Humanities scholarship, while Humanities scholars expressed the belief that they should further develop their technical understanding so as to better communicate with collaborators. These results are arguably a product of the sample group, which was comprised of scholars who considered themselves as working within the Digital Humanities. However, what it demonstrates nonetheless is that there is a genuine desire for collaboration within the discipline, and that this desire is supported by an awareness of differing expertise, requirements and mindsets.



2.4 Leadership

It is clear from the activities of respondents that digital projects are being managed by Humanities scholars. Of those respondents that stated they had worked on a digital initiative, the majority did so in a project management capacity, as illustrated in Figure 4.



This is a positive finding, as it suggests that technology is being used to support the agendas of the Arts and Humanities, rather than dictating what such agendas might be.

References

1. stephenramsay.us/text/2011/01/08/whos-in-and-whos-out/

Measuring the style of chick lit and literature

Jautze, Kim Johanna

kim.jautze@huygens.knaw.nl

Huygens Institute for the History of the Netherlands

1. Introduction

Novelistic genres have certain formal conventions. By close reading novels of various genres, structural features such as the differences in theme, motifs and plot can be observed. According to Jockers (2013), there seem to be stylistic differences, caused by the author's linguistic choices, between genres as well.

In this paper I examine the stylistic fingerprints of chick lit and literature in terms of the most frequent words. The focus on the styles of these particular genres relates to the overarching question of my research: does the perception of literariness (e.g. the dichotomy between "highbrow" and "lowbrow" genres) correlate with certain linguistic characteristics, or with the degree of variety of these in the style? The aim of this pilot study is to explore how well successful stylometric methods could be applied as a starting point for such a comprehensive question.

2. Background

Styliometrists who study linguistic patterns in fiction typically focus on classification tasks, e.g. authorship attribution or text genre detection. The latter studies usually examine how well certain texts can be identified into pre-defined classes; for instance *editorials*, *newspapers* and *literature* (Stamatatos et al., 2000). Styliometric studies of novelistic genres (e.g. Louwes et al., 2008; Ashok et al., 2013) seem to be scarce. The most extensive study is performed by Matthew Jockers (2013) and his colleagues at the Stanford Literary Lab (Allison et al., 2011). They examine to what extent formal conventions can be detected at the level of the high-frequency function words. Jockers (2013) concludes that genres to a certain degree have measurable linguistic fingerprints, and that linguistic decisions of the authors seem to be dependent upon their genre choices. An interesting next step would be to analyze and interpret these fingerprints, as has been done for authorial markers by Burrows and Craig (2012). In this paper I adopt their approach to examine the stylistic differences between chick lit and literature.

A previous study into the deep syntactic structures of the two genres shows that literary authors tend to use more complex sentences and employ a more descriptive language, whereas chick lit to a greater extent resembles colloquial speech (Jautze et al., 2013). In the current study I complement this syntactic characterization with a stylometric analysis of the function words. These words relate to syntactic structure because they add grammatical information by organizing and connecting the content words. The question arises if the most frequent words (MFWS) differentiate between the two genres. And if so, do these linguistic patterns give more insight into the two genre styles?

3. Materials and method

According to Jockers (2013) it is hard to distinguish linguistic fingerprints that are related to the time of writing from actual genre signals. This means that when one wants to examine genre fingerprints, the chronology factor must be ruled out as far as possible. My corpus therefore comprises 32 original Dutch novels (16 literary and 16 chick-lit) of the last two decades.

In order to computationally examine the style of the two genres I start with the stylometric approach to search for the style markers. The stylo package in R compiles a word

frequency list for the entire corpus (Eder et al., 2013). Then, I want to explore the language patterns to characterize the two genre styles. Egbert (2012) argues that lexicogrammatical features can be captured in three dimensions of discourse presentation. Two of these dimensions I will adopt in this analysis in order to analyze the linguistic patterns: (i) *description* versus *thought representation* and (ii) *dialogue* versus *narrative*.

4. Results

Figure 1 shows that the titles per author as well as the two genres cluster together (the abbreviations indicate the pre-defined genres). According to Jockers (2013), men and women tend to use different (function) words, so one might have expected the female literary authors in my corpus to cluster together with the female chick-lit writers. My results indicate the opposite, which suggest that (in this corpus) genre signals precede gender signals. It is striking however that the chick-lit writers are more grouped together than the literary authors. This indicates that there is more variation within the literary writing styles.

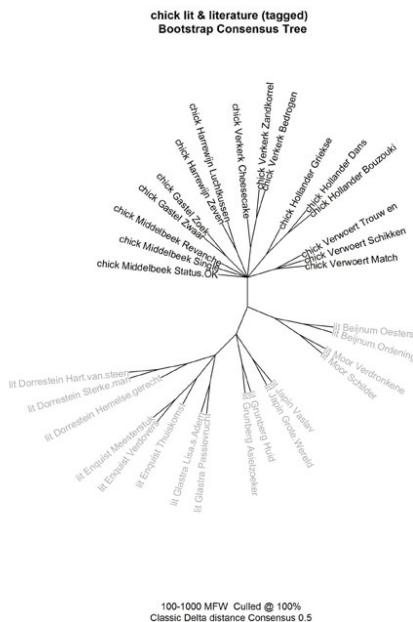


Fig. 1: A Bootstrap Consensus Tree showing average similarity of texts based on the frequencies of 100-1000 MFW.

In order to examine linguistic patterns behind this genre-distinction, the word frequencies are analyzed. A Principal Components Analysis uses the MFWs as variables according to which the texts are correlated in a matrix. Figure 2 shows that the 100 MFWs map the genres in separate areas of the graph, except for chick-lit writer Wilma Hollander. She sides with the literary authors.

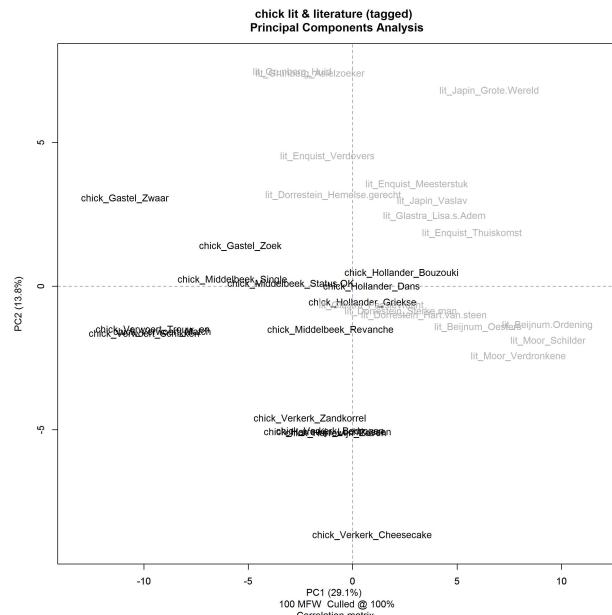


Fig. 2: A PCA showing the plotting of texts based on the weightings of 100 MFW.

The two components of the PCA together account for 43.7% of the variation between the novels. The word-variables have their own weightings for each component according to which the texts are scored in the matrix (e.g. Figure 3). In the previous study by Jautze et al. (2013) the novels were parsed with the Alpino parser (Bouma et al., 2001). The parts-of- speech tags made it possible to separate homographic word forms, as in *zijn* ('his') and *zijn* ('are').

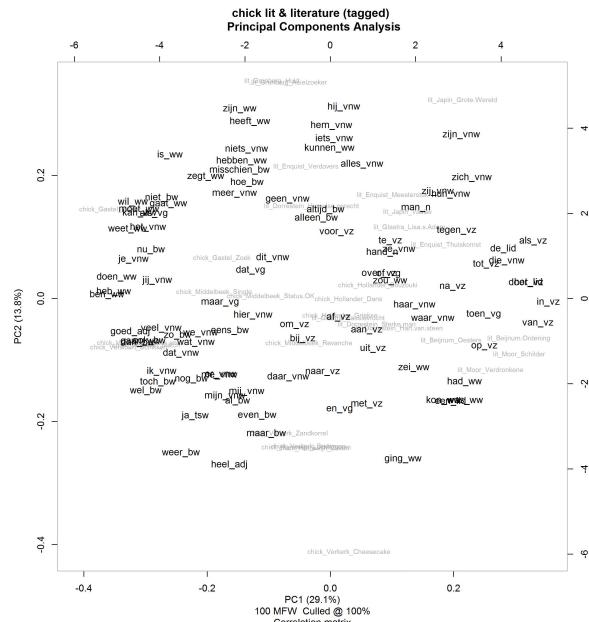


Fig. 3: A PCA showing the plotting of 100 word-variable weightings.

POS tags	Translation tags
Vnw	Pronoun
Ww	Verb
Bw	Adverb
Vg	Conjunction
Vz	Preposition
Lid	Determiner
Adj	Adjective
N	Noun
Tsw	Interjection

With regard to Egbert's dimensions, it can be argued that the literary authors employ more *descriptions* and *narratives*, whereas in chick lit more *thought representations* and *dialogues* are used. Indicative for the descriptive dimension is the high amount of prepositions, the use of determiners and the demonstrative *die* ('that'). Prepositions express spatial or temporal relations between subjects and/or objects, and therefore are used for detailed-oriented description. Along with the use of determiners and demonstratives, this indicates that the literary authors use relatively more nouns. These findings can be underlined by the results of Jautze et al. (2013), that show that noun phrases and prepositional phrases occur more frequently in the literary books than in the chick-lit novels of our corpus.

Other frequent "literary" function words in the PCA are third person pronouns such as *hij* and *hun* ('he' and 'their/ them'), indefinite pronouns such as *iets* and *alles* ('something' and 'everything') and verbs in the past tense. According to Egbert (2012), these linguistic features belong to the narrative dimension. Especially the past tense verbs indicate that the literary narrators describe events. The chick-lit writers on the other hand, employ more present tense verbs, and first and second person pronouns such as *ik*, *mij* and *jij* ('I', 'me' and 'you'). These, as well as the demonstratives *dat* and *daar* ('that' and 'there'), are argued to be indicative for the dialogue dimension.

Moreover, at the chick-lit side of the plot a lot of words are mapped that relate to the dimension of thought representation. Function words like the mental verb *weet* ('know'), the indefinite pronoun *veel* ('many'), the affective adjectives *heel* and *goed* ('very' and 'good'), the possibility modal *kan* ('can') and the likelihood adverb *misschien* ('maybe') offer insight into the narrator's or character's psyche. The chick-lit authors also employ certain adverbs (*maar*, *toch*) that can cause an emphatic effect. It could be compared with 'there are *only* seven'. It shows a character's or narrator's involvement, and it belongs to a more colloquial language register.

5. Conclusion

The results of this paper show that stylometric analysis can be used in stylistic research of literary genres. The linguistic patterns detected in this small corpus suggest that the literary authors have a more detail-orientated descriptive style when compared to the chick-lit style, which tends to be more informal and involved. The preliminary results offer a variety of clues to further research. In the next stage I would like to explore for instance word n-grams and parts of speech in a corpus that is expanded with several other "highbrow" and "lowbrow" genres.

Acknowledgements

I am grateful to my supervisor professor Karina van Dalen-Oskam and to Corina Koolen and Andreas van Cranenburgh for reading my drafts, and to Andreas for assisting with tagging the parts of speech in the novels.

References

Chick-lit novels humorously address the challenges of young urban female protagonists.

his study is part of The Riddle of Literary Quality Project. In this project we explore the assumption that formal characteristics play a role in the aesthetic appreciation of novels. Cf. literaryquality.huygens.knaw.nl

This is the same corpus as has been studied in Jautze et al. (2013). Ideally, female and male writers should be equally represented. But since the chick-lit novels were all written by women, this was not possible.

This Bootstrap Consensus Tree is a mean of ten cluster analyses, varying from 100-1000 MFWs with an increment of 100. The corpus is culled at 100%, which means that words that are unique for individual texts are removed.

Allison, S., Heuser, R., Jockers, M., Moretti, F. and Witmore, M. (2011). *Quantitative Formalism: An Experiment*. In Pamphlets of the Literary Lab 1, litlab.stanford.edu/LiteraryLabPamphlet1.pdf , (accessed on 24 October 2013).

Ashok, V.G. , Feng, S., and Choi, Y. (2013). *Success with Style: Using Writing Style to Predict the Success of Novels*. In Empirical Methods on Natural Language Processing, Seattle. 1753-1764, aclweb.org/anthology/D/D13/D13-1181.pdf , (accessed on 28 October 2013).

Bourma, G., Van Noord, G. and Malouf, R. (2001). *Alpino: Wide-coverage computational analysis of Dutch*. In Language and Computers, 37 (1). 45–59.

Burrows, J. and Craig, H. (2012). *Authors and Characters*. In English Studies 93 (3). 292-309.

Eder, M., Kestemont, M. and Rybicki, J. (2013). *Stylometry with R: a suite of tools*. In Digital Humanities 2013: Conference Abstracts. Lincoln (NE). 487-489.

Egbert, J. (2012). *Style in nineteenth century fiction*. A Multi-Dimensional analysis. In Scientific Study of Literature 2 (2). 167-198.

Jautze, K., Koolen, C., Van Cranenburgh, A. and De Jong, H. (2013). *From high heels to weed attics: a syntactic investigation of chick lit and literature*. In Proceedings of the Workshop on Computational Linguistics for Literature. Atlanta (GA). 72-81, aclweb.org/anthology//W/W13/W13-1410.pdf .

Jockers, M. L. (2013). Macroanalysis: Digital Methods and Literary History. Illinois: University of Illinois Press.

Louwerse, M., Benesh, M.N. and Zhang, B. (2008). *Computationally discriminating literary from non-literary texts*. In: Zyngier, S., Bortolussi, M., Chesnokova, A. & Auracher, J. (Eds.), Directions in Empirical Literary Studies. Linguistic Approaches to Literature 5, Amsterdam. 175-191.

Stamatatos, E., Fakotakis, N., & Kokkinakis, G. (2000). *Text Genre Detection Using Common Word Frequencies*. In Proceedings of the 18th conference on Computational linguistics (2). Stroudsburg (PA). 808–814.

All databases are created equal: building profiles for database standards and interoperability in the Humanities

Johnson, Ian R.

University of Sydney, Australia

In this paper I will discuss the development of standard database profiles which shortcut the process of building complex interlinked Humanities databases. Although there is a risk of creating restrictive uniformity which stifles creativity, I argue that the benefits of well-structured starting points far outweigh the drawbacks, both in terms of immediate productivity and the avoidance of less than optimal structures. I will use examples from two national infrastructure projects to illustrate the use of database profiles to provide a database-on-demand service which integrates readily into an aggregated search of cultural datasets and specific user needs.

Over the past few years I have worked with a number of projects - notably historical and archaeological projects - to

model their database needs in Heurist (HeuristScholar.org). From these models I have been able to generalise a set of requirements for commonly encountered entity types, and their interrelationships. While there is significant variation in the specific fields required to describe entities (notably the degree of detail required), there are a range of common entity types such as - for historic data - person, organisation, building, ship, voyage, epidemic, performance, venue, work and other bibliographic types, along with repeatedly used descriptors (fields), which are common to a range of projects. Furthermore, there are a range of relationships between entities, including familial relationships, roles, event relationships and bibliographic relationships, which are widely shared.

The role of a database profile is first to shortcut the low level and time-consuming task of defining a set of commonly used descriptors. For example, the description of a person - one of the most widely used and easily standardised entities - will commonly require some or all of family name, given names, sex, date of birth, various forms of address and so forth, with a bifurcation between contemporary individuals (eg. participants, with an email address and phone numbers) and historical individuals (eg. slaves with national or racial origins and date of death). Most of this is not demanding, but it is still common to see poorly structured descriptors, (such as text for categorisation fields or loosely structured coordinate data in place of geometries), with little or no metadata, beyond a field or column name, identifying the nature of the data recorded. A pre-populated entity description ensures good, clear, documented descriptors, which will rarely be perturbed by, or stand in the way of, individual customisation.

More critically, there are several alternative ways of building relationships between records, ranging from simple pointers or foreign keys to typed relationships with annotation and temporal range. Relationships may be constrained with specific cardinality eg. a person can only have two biological parents, but uncertainty or cultural perspectives can complicate such simple rules. The particular solutions adopted in handling relationships can have important ramifications down the line when it comes to searching, analysis and presentation, but these ramifications lie outside the experience of a researcher lacking a data modelling background. Even with such a background, it can be hard to identify the optimum solution without some trial and error.

Heurist has been incorporated into two national infrastructure projects - HuNI (Humanities Networked Infrastructure <http://huni.net.au>) and FAIMS (Federated Archaeological Information Management System <http://fedarch.org>). In both cases, Heurist provides a database-on-demand service, allowing researchers to build their own databases on the NeCTAR Research Cloud, or on their own servers (University or commercial IP), without recourse to technical assistance. For each of these projects we have therefore developed an initial database profile which reflects the needs of the community, drawing on both the ontologies developed by these infrastructure projects and on our experience in developing practical data models for numerous individual projects.

The HuNI project has put significant effort into establishing an ontology (<http://wiki.huni.net.au/display/DS/Input+Data+Sources+Model>) to provide interoperable search across the 23 cultural datasets aggregated in the system. The Heurist HuNI profile reflects this ontology and incorporates a mapping between human-friendly descriptive terms (which researchers can modify to suit their needs), and the standardised terms required by HuNI. This allows any database based on this template to immediately generate XML which can be harvested and searched within the HuNI framework.

The FAIMS project has also identified shared units of description, although it has not generated a formal ontology. The FAIMS profile reflects this understanding through a common set of archaeological entities such as project, transect, survey unit, site, trench, phase, layer, context, feature, find, sample etc., as well as relationships such as roles, stratigraphic and contextual relationships. However, in archaeology, recording systems used are often influenced by national and state legislation, as well as differing traditions of survey and excavation methodology, so FAIMS is also building a Heurist

database containing multiple alternative recording systems deployed as a Community Server. Structural elements - entity types with all their associated fields, term lists and relationships - can be imported selectively into a database created with the FAIMS profile while retaining field and term mappings across different systems; interoperability is thus not legislated by fixed structures, but encouraged by reuse.

Local voices, worldwide conversations: ethnographic methodologies as a route to understanding meaning and value of niche local digital cultural heritage resources.

Johnston, Penny

University College Cork, Ireland

Oral history materials are increasingly disseminated and accessed online. Audio, the primary "document" of the discipline, is now easy to publish in web-based media. It is no exaggeration to say that digital technologies have transformed the practice of oral history. At the same time, the digital has opened up new areas of debate, including the question of assessing the value of online oral history initiatives, a theme that is increasingly common when examining digital cultural heritage resources in general (see, for example, Tanner 2012). This poster presents an ongoing research project that examines the appropriateness of local resources for global audiences, and how the value of these niche projects can be assessed in a meaningful way. This will be combined with a chance to try out an online oral history project, and to engage in discussions with, and elicit feedback from, the international digital humanities community.

The focus of this research is a digital oral history case study from Ireland, the Cork Memory Map (www.corkmemorymap.org). This is one of the online initiatives created by the Cork Folklore Project, a community-based oral history organisation that was set up as a collaboration between university and community and has been collecting everyday stories of life in Cork city since 1996. As a long-standing collection and archiving centre, the Cork Folklore Project has well-established methodologies and understandings that facilitate its work within its host community, with living "subjects", and all the attendant issues that arise in such situations, particularly those associated with consent, copyright and duty of care towards participants.

Embedded audio and images in the Cork Memory Map help to explore the narratives and stories associated with the landscape and "culturescape" of the Cork city (see O'Carroll 2011, 184), exploring meaning-making, identity, cultural heritage and local place-making through memory and narrative. Projects such as the Cork Memory Map have a strong local appeal, but how are these received in a global forum such as the web?

This ongoing research project uses empirical techniques (such as website and social media metrics) and ethnographic methodologies to investigate the questions of value and meaning in digital cultural heritage resources. Some preliminary results will be presented in the poster. Qualitative research will involve conducting ethnographic interviews with many stakeholders in the Cork Memory Map, including the contributors to the map (researchers and interviewees), academics and historians interested in local and oral histories, people from Cork (or with a link to the city), as well as the wider global audience that it is now possible to reach through online dissemination. The aim is to interview as wide a variety of users and stakeholders as possible.

At Digital Humanities 2014 the poster presenting this research will be more than just a poster – the aim is interactivity, with the poster and accompanying elements of the display (laptop displaying the Cork Memory Map and a scrapbook of ethnographic field notes) designed not only to showcase but also to elicit responses from an international community of digital humanities researchers. These opportunities for interaction and discussion will be used as material to be incorporated into the ethnographic research, and to explore potential new avenues of research. The poster and interactions will therefore be an attempt to integrate an academic conference presentation into an ongoing process of research.

References

- O'Carroll, C. (2011). *The Cork Memory Map*. Béascna Journal of Folklore and Ethnology, 7, 184–188.
- Tanner, S. (2012). *Measuring the impact of digital resources: the balanced value impact model*. Department of Digital Humanities, King's College London. Retrieved from www.kcds.kcl.ac.uk/innovation/impact.html

Data Criticism: General Framework for the Quantitative Interpretation of Non-Textual Sources

Kitamoto, Asanobu

kitamoto@nii.ac.jp

National Institute of Informatics

Nishimura, Yoko

nishimura@toyo-bunko.or.jp

Toyo Bunko

1. Introduction

The usage of non-textual sources for historical studies, in addition to conventional textual sources, has a potential to expand our historical knowledge, but non-textual sources have been used less frequently than textual sources, and framework for the critical usage of non-textual sources is under developed. We therefore are working on a framework called "data criticism" to develop methodological commons for the quantitative evaluation of quality and value of non-textual sources (= data sources), such as maps, photographs and illustrations. Here, in the same sense with textual sources, we should not forget critical evaluation in spite of its numerical or graphical appearance that may be misunderstood as "the objective representation of facts." The fundamental requirement of using sources for historical studies is that we should critically "read" sources to reduce uncertainty, biases, and other factors that affect the quality of facts derived from sources.

2. Data Criticism

We have been working on Digital Silk Road project (dsr.nii.ac.jp) since 2001, starting from the digitization of historical sources to the analysis of digitized materials for historical studies, especially in the area of Silk Road. We realized that not only textual sources, but also non-textual sources have the challenge of proper interpretation because many types of errors were found. Maps had errors due to technical limitation at the era of making maps, and photograph captions had errors due to misunderstanding or different conceptualization. While working on fixing these problems, we realized that this task is exactly the non-textual version of textual criticism (text critique) which is at the core of historians' research. We therefore propose the concept of 'data criticism' both to clarify the role of the task in the whole process of historical research, and to raise attention to the importance

of computational tools, namely algorithms and databases, for processing data. However, the development of computational tools is required for humanists to pursue the potential of data criticism because off-the-shelf tools are usually not available. Hence the contribution of the paper is to demonstrate with case studies how maps, photographs and illustrations can be critically interpreted and used as historical evidences, and also to provide digital tools to support humanities perform this task.

Data criticism has relationships with a few similar concepts. Data criticism on maps, or what we can call map criticism, has much in common with methodologies developed in cartography and geography (for example, reference), but data criticism focuses more on the integration of multiple spatial and visual sources, or also textual sources such as placenames, to cross-check the interpretation of historical landscape. Data criticism can also be characterized as a quantitative approach in comparison to a qualitative approach used previously by historians to "read" evidences from paintings or illustrations based on human interpretations. Here the word 'interpretation' needs to be clarified because historical studies deal with two levels of interpretation, namely the evidence level and the history level. In particular, the latter focuses on answering historical research questions, but this answer depends more on the cognitive process of how historians construct the history. The word 'capta' is suggested to clarify the constructivism context of history in comparison to the word 'data' that sounds more objective and observer independent. In this sense, our interest is in the evidence level (data), not in the history level (capta). The goal of data criticism is to provide more reliable evidences from data sources through quantitative and integrated procedures using computer algorithms.

3. Case Studies

We introduce data criticism with three case studies from the achievement of our project, namely Digital Silk Road Project.

The first case study is about a map of Silk Road made by Aurel Stein in the beginning of 20th century. This map is still considered as the authoritative reference of Silk Road studies, but the map has a mysterious problem; some of the ruins recorded on the map cannot be found at respective geographic coordinates. A typical interpretation of this mystery explains that those ruins were destroyed or disappeared since his visit of about 100 years ago. We discovered, however, from archaeological survey that those "missing ruins" are still there, but at the location suggested by the map plus the error of the map estimated from ground control points (GCP) given on the map for geometric correction. This 'place-matching' result is also supported by evidences from non-textual sources such as photographs and illustrations, because they provide information about ruin's 2D or 3D structures which has less probability of coincidence. After this matching, we also realized that the comparison of ruins' name, or 'name matching' does not work, because a string comparison of names cannot overcome the problem of linguistic difference between endonyms and exonyms, or different naming conventions due to arbitrary conceptualizations. This case study suggests that the integration of multiple non-textual sources can provide more reliable evidences and new interpretations of historical facts.

The second case study is about a map of Gaochang made by Albert Grünwedel in the beginning of 20th century. This is an important map because artifacts excavated from the site were recorded by ruin symbols on the map, but the accuracy of the map has been considered as untrusted because the map seems to be a sketch with significant distortion. We hypothesized that this is a topological map, just like a subway map, designed for navigation purposes by preserving topological relationships such as connection and intersection of geographical features. We criticized the topological structure of the map, and finally identified most of the ruins recorded on the map. Photographs are again used to support evidences about the identity of ruins. For this purpose, we developed a web-based tool, 'mappining' for what we call 'interactive georeferencing.' This tool is designed to realize on-the-fly registration of two maps at the focal point specified interactively by a user. This method can avoid significant distortion of the

topological map in comparison to geometric correction of the entire map, so historians can keep the original shape of the map, while taking advantage of approximate place matching in the neighborhood of the focal point.

The third case study had the highest technical challenge, because the target map, Complete Map of Peking, Qianlong Period, made about 250 years ago, is a huge map having 29 billion pixels and being separated into 203 sheets. Massive geometric correction of this map required 1800 ground control points and 500 control lines. The uniqueness of this geometric correction is that we used not only control points but also control lines to maximally preserve the linear features of streets in Beijing through geometric correction. This was the first time in the world to obtain the digital version of the fully-connected and geometrically-corrected map.

After geometric correction, we realized that some parts of the map could not be matched well with the current map. This problem was known to some historians, but they could not explain the reason, so they simply judged that this map cannot be trusted. We found, however, that this problem can be explained by the erroneous reconstruction of the broken map at some point in history. We discovered the exchange of sheets from left to right or vice versa, which cannot be caused by mistakes in the digitization process, because some exchange can be observed within a single map sheet. This insight was obtained only after we had an entire picture of the map, and historians in the past could not notice this problem as long as they study the map sheet by sheet. This indicates that human interpretation in a microscopic scale has limitation in understanding the macroscopic issues, and there has been technical barriers to realize this viewpoint. Data criticism tools have potential to break this barrier and may lead to new discoveries based on new interpretations from new viewpoints.

These case studies suggest that the proper treatment of non-textual sources needs an integrated approach using not only maps but also photographs and other spatial and visual sources. Photographs should be interpreted in a three-dimensional historical landscape to identify the location and direction of the photograph. This interpretation should be supported by textual criticism of photograph captions that may also be affected by errors or misconceptions.

4. Conclusion

In short, data criticism is not about making historical Geographic Information Systems (GIS) that maps historical facts into digital space and analyze them, but about making quantitative and integrated digital tools to analyze, enhance and re-discover the value of historical sources. We plan to generalize our framework to establish the field of "data criticism" so that we can accumulate historical evidences not only from textual sources but also from data sources. The key is to develop easy-to-use digital tools for humanists to be used by themselves. This is where digital humanities comes in; where team work between computer scientists and humanists can lead to a breakthrough. The final goal is to establish data criticism as a new research field that has its own methodological commons and framework to combine appropriate tools to answer historical research questions.

References

- Ono, K., KITAMOTO, A., Onishi, M., Andaroodi, E., Nishimura, Y., Matini, M.R. (2008), *Memory of the Silk Road - The Digital Silk Road Project-*, Virtual Systems and Multimedia (VSMM), Vol. Project Papers, pp. 437-444
- Gregory, I.N., Healey, R.G. (2007), *Historical GIS: Structuring, mapping and analyzing geographies of the past*, Progress in Human Geography, Vol. 31, No. 5, pp. 638-653.
- Drucker, J. (2011), *Humanities Approaches to Graphical Display*, Digital Humanities Quarterly, Vol. 5, No. 1
- Silk Road Maps*, dsr.nii.ac.jp/geography/
- Kitamoto, A., and Nishimura, Y. (2009), *Geometric correction of measured historical maps with a pixel-oriented and geobrowser-friendly framework*, Proceedings of the

22nd International Symposium on Digital Documentation, Interpretation & Presentation of Cultural Heritage (CIPA)
Mappinning - Interactive Georeferencing by Pinning Old Maps, dsr.nii.ac.jp/digital-maps/mappinning/
Digital Maps of Old Beijing, dsr.nii.ac.jp/beijing-maps/
 Gregory, I.N., Healey, R.G. (2007), *Historical GIS: Structuring, mapping and analyzing geographies of the past*, Progress in Human Geography, Vol. 31, No. 5, pp. 638-653.

Shedding Light on Dickens' Style Through Representativeness & Distinctiveness

Klaussner, Carmen

klaussnc@tcd.ie
 Trinity College Dublin

Nerbonne, John

j.nerbonne@rug.nl
 University of Groningen

Çöltekin, Çağrı

c.coltekin@rug.nl
 University of Groningen

The present work is an exploration into Dickens' style using the statistical method of *Representativeness & Distinctiveness*¹ to detect words that Dickens either particularly preferred or avoided compared with other writers of his time. It takes as a starting point research reported on in DigitalHumanities 2012², where Tabata used the classification algorithm *Random Forests*³ to determine words able to distinguish Dickens' works from both contemporary author Wilkie Collins and a larger set of authors from the 18th and 19th century.

Representativeness & Distinctiveness was originally conceived in the realm of dialectometry, where it has been shown to detect lexical items distinguishing different dialectal areas. In the context of stylometry, the method detects elements for which an author is consistent throughout his own works while also separating him from others. Considering, for instance, a comparison between Dickens and fellow writer Collins on word features using a couple of novels of each writer, one first determines Dickens' **representative** terms, i.e. those words which he uses consistently either frequently or infrequently over his works. In order to arrive at a combined measure, one then favours those representative terms of Dickens that Collins uses either inconsistently or consistently but with a different frequency over his novels. The remaining group of words are considered to be Dickens' **representative** and **distinctive** terms when compared with Collins. Since the analysis is directional, the degree of Representativeness of individual features being different with respect to each author, comparisons are made twice - once from Dickens' to Collins' set and once from Collins' to Dickens' set. This returns two individual author profiles, where features occurring in both profiles are also consistent for both writers.

Thus, Representativeness & Distinctiveness bears similarities with both Burrow's *Delta*⁴ and *Zeta*⁵ in so far as favouring consistent terms that are irregular in the opposing author's set. Additionally, it is also similar to *Zeta* in being dependent on the other set for the selection of distinctive terms out of the representative ones. However, rather than preselecting words according to different frequency strata, it is used here on the first 5000 most frequent ones.

We compare our results on "Dickens vs. Collins" and "Dickens vs. World" to the earlier study using Random Forests (RF) classification that was also done on the basis of frequency comparisons. RF is a machine-learning technique based on ensemble learning from a large number of decision trees, hence "forests". Each tree is trained on a different subset of the training data and subsequently evaluated on the remainder. At each node a different subset of the total features are considered and selected according to the best split. Individual features'

importance is averaged over all trees and similar to our method there is a measure of how useful a particular feature is for classification. Representativeness & Distinctiveness and Random Forests are conceptually similar in that the document space is considered as a set of smaller comparisons between documents and distinguishing features are chosen accordingly. While RF presents the complete process of classification and evaluation, Representativeness & Distinctiveness, although less straightforward to evaluate, might have advantages in terms of interpretability. Since a gold-standard indicating the most prominent stylistic features of an author is generally not available, we evaluate author profiles over different iterations in cross-validation by testing how well the selected words are able to separate authors in clustering. For this purpose, we retain the shared words of both profiles, since the values for terms in only one author profile might not be consistent in the other author's documents. The next step is then to cluster all documents based on the relative frequency for those shared terms. Based on the ideal clustering result into the two author groups and the present iteration's clustering result, we compute the *Adjusted Rand Index*⁶. The evaluation technique proposed here is intended to measure classification ability of the author profiles and thus also appropriateness of the method responsible for choosing them. Applying it to our representative and distinctive profiles of Dickens and his contemporaries indicates a high degree of separation ability in clustering. Despite using a different method to determine the author profiles, there is a fair overlap of items with the earlier study using Random Forests, which would strengthen the assumption of them being genuine stylistic elements of Dickens and his peers.

References

1. Jelena Prokić, Çağrı Çöltekin, and John Nerbonne (2012). *Detecting shibboleths*. In: Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH. EACL 2012. Avignon, France: Association for Computational Linguistics, pp. 72–80.
2. Tomoji Tabata. (2012). *Approaching Dickens' Style through Random Forests*. In: Proceedings of the Digital Humanities. Hamburg, Germany.
3. Leo Breiman (2001). *Random Forests*. In: Machine Learning, pp. 5–32.
4. Burrows, J. (2002). 'Delta': a measure of stylistic difference and a guide to likely authorship, Literary and Linguistic Computing, 17 (3). 267–87.
5. Burrows, J. (2007). *All the Way Through: Testing for Authorship in Different Frequency Strata*. LLC 22: 27–47.
6. Lawrence Hubert and Phipps Arabie (1985). Comparing partitions. In: Journal of Classification 2.1 (Dec. 1985), pp. 193–218.

Finding Inexact Quotations Within a Tibetan Buddhist Corpus

Klein, Benjamin Eliot
Tel Aviv University, Israel

Dershowitz, Nachum
Tel Aviv University, Israel

Lior, Wolf
Tel Aviv University, Israel

Almogi, Orna
Universität Hamburg, Germany

Wangchuk, Dorji
Universität Hamburg, Germany

Introduction

One thing that literary scholars routinely look for – regardless of the specific field – is textual citations, where one work quotes or paraphrases another work. In historical works, even quotations are frequently quite inexact. To complicate matters further, there is often no clear indication that a passage is being quoted, let alone which work is being cited. It is, therefore, only natural to use algorithmic tools to search for such occurrences in texts and present the results to scholars for consideration.

One such corpus is the Tibetan Buddhist canon. Altogether, there are more than 300 volumes, averaging about 800 pages (400 folios) of 200 words each. In addition to the canon, there are many other important collections in the Tibetan Buddhist literary corpus. And, of course, there are many Buddhist corpora in other languages. Scholarly editions are still wanting for many of these works, so we have set out to design computerized tools to help deal with the masses of data.

The most relevant previous work is by Prasad and Rao [2] who search for citations within Sanskrit texts. They break the text into units (lines, say) and then compare each potential citation with each unit in the cited corpus (Smith-Waterman-Gotoh), using approximate match. To constrain the search, they first sort the units, so they only need compare units that begin similarly. We approach the problem of reducing the complexity of the search differently. We borrowed an algorithm designed for finding all (sufficiently long) approximate subsequence matches in genomic data and adapted it for finding common approximate subtexts between two large corpora. This involved parallelizing the algorithm, adding some simple preprocessing, and some less trivial post-processing.

It is important to distinguish between various string-matching tasks. Given two or more passages known to be similar, or several recensions of the same work, one can seek the best alignment. The task we address of finding all, short or long, approximately similar texts that appear at arbitrary locations within large corpora is vastly different from the alignment tasks Juxta (available at <http://www.juxtasoftware.org/>) and CollateX (<http://collatex.net/>) help solve. In order to address the reviewers' concerns we tried these tools. Juxtas results on our texts were completely irrelevant; CollateX, after running for a long time, produced an empty output.

2 The Corpus

In this exploratory work, we compared two major Buddhist texts, transliterated into Latin characters from the Tibetan, the *Sūtrasamuccaya* and the *Sikṣāsamuccaya*. Each is over 150,000 words long.

The *Sūtrasamuccaya*. The "Compendium of [Citations from Mahāyāna] Sūtras" is a compilation ascribed to the famous Nāgārjuna (2nd century ce). The text has survived only in its Tibetan (P5330, D3934) and Chinese translations. It is an important source for early Mahāyāna sūtras, and it is invaluable for the study of the early phase of Mahāyāna. The Tibetan translation was done by the Tibetan famous translator Ye-shes-sde (8th century), in collaboration with Jinamitra and Silendrabodhi.

The *Sikṣāsamuccaya*. The "Compendium of Teachings" (i.e. citations from mostly early Mahāyāna sūtras) was compiled by the famed Indian scholar Sāntideva (7th c.). It was translated into Tibetan by the same Tibetan translator Ye-shes-sde (8th c.), in collaboration with Jinamitra and Danśīla. Later on the translation was revised by the Tibetan translator Blo Idan shes rab (1059–1109) in collaboration with the Kāśmirian scholar Tilakakalaśa.

3 Method

We designed an algorithm to solve the problem of finding local regions with high similarity in the two texts. The main workhorse is an efficient algorithm for solving the "threshold all against all" variant of the problem, based on that of Barsky et al. [1]. It finds all maximal substrings of the text of some minimal length $L_0 = 60$, with an edit distance between them bounded by some given value $k = 10$.

We worked with transliterations of the Tibetan texts. The texts contain multiple spaces, line breaks, page numbers, punctuation marks and the like. In a preprocessing step, we clean the texts and remove all such.

The texts we are using are very long and therefore we created a parallel version, running on large overlapping chunks of length $l = 25000$ on a cluster of processor cores. After collecting all the results, some post-processing steps are required in order to build a non-redundant and meaningful collection of local regions with high similarity. Splitting increases the quantity of overlapping results. In addition, some results are very near to each other and should be merged into a longer match. We address these issues by uniting every pair of overlapping or nearby results.

The second problem arises when we have a meaningful result with length that is smaller than the threshold and with a very small edit distance. In this scenario, the algorithm extends the result in order for the minimal length constraint to be satisfied, resulting with a less meaningful result. We solve this issue by applying local alignment on each result, which removes these uninformative extensions.

From the final collection of matched substrings we get two main outcomes; the trivial one is finding the local regions of high similarity in the two texts. We built a designated interface that has tools for investigating the matches. It presents the two texts side by side and a list of all the matches in descending order of their edit distance. Using it, one can focus on a specific match, and see the relevant substrings in both texts. The substrings are also presented in another window, aligned to each other for convenient comparison. Additionally, by selecting a substring in one of the texts, one can see all the matches that overlap with the selection. See the screenshot in Fig. 3.

The second outcome is computing statistics on all the results. This requires us to carefully align each result, as the quality of the statistics depends on the alignment of each single word. The alignment is done by a variant of global alignment that penalizes gaps that occur between words (or at one of the ends of the string) differently from gaps that occur within a word. This simple alignment allows us to derive meaningful statistics on differences between collections.

4 Results

Overall, 2514 matches were found between the texts. These matches cover a significant fraction of the texts and 9.15% of the Sūtrasamuccaya as well as 10.85% of the Sūtrasamuccaya as 10.85% of the *Sikṣāmuccaya* of regions for which at least one match in the other text was found. Some of the matches are quite long, as can be seen in the histogram in Fig. 1. Example matches, as exported to files by the developed research tool, are presented in Fig. 2.

Sample matches could be verified and found to be correct. In some cases, however, they needed to be extended to regions before or after the marked texts. Many of the omissions and substitutions were not surprising, such as *cing* --> *zhing*, *bting* --> *gding*, and *po wang* --> *pi bang*. Some of the variations seem to be simple typos, either by the scribe of the original or simply during the digital transliteration. Typical transcription errors include substitutions between b and p, or ng and d, which appear similar in Tibetan. In some cases, making the distinction whether a variant stems from an accidental typo or is in fact a substantive variant is not clear. For example, *gtor* means to scatter and *'thor*, which was correctly matched is some of the quotations, means to be scattered. Table 1 exhibits the “confusion matrix”, i.e., the most common letter substitutions between the two texts.

5 Discussion

Two well-known and studied works were chosen as a trial case for our first experiment, namely, transliterated texts of the two Buddhist works that were translated into Tibetan in the 8th century. Due to the large number of shared citations found in these works, they made for a good trial case for algorithmically locating matches. But because these are two

different anthologies (i.e., not two versions of the one and the same work), they are different enough to provide sufficiently many instances of approximate matches and discrepancies.

Besides the value of the citations themselves, in the absence of a critical edition of one of the works, statistics regarding the types of variations (particularly in the usage of particles) hint at the nature of the editing done by the 11th-century revisers. Through the statistics, scholars may be able to learn about some stylistic differences and editorial practices. Continuing this line of work, we may even have a case where through a careful analysis of the differences one could become aware of some philosophical differences and developments.

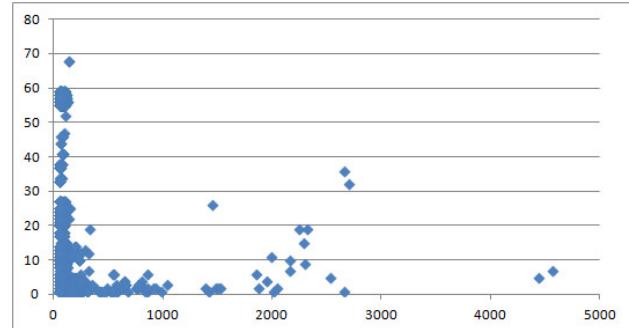


Fig. 1: Histogram depicting the count (y-axis) of matching texts of a certain length (x-axis) [the length in Sūtrasamuccaya]. As can be seen, while most of the matches are for texts of up to 200 characters (median=61), there are also matches for texts of a few thousand characters. Note that some texts (including long ones) have multiple matches.

It has been argued before that the so-called “revisions” often involve only very minor and unimportant changes, and indeed some of the revisers had often been accused by some Tibetans of plagiarism. The difficulty, however, is that in most cases we only have access to the revised version(s), and thus cannot compare the revision with the initial translation. Our alignment and statistical tools can help scholars trace and re-evaluate this phenomenon.

Lastly, going beyond direct matches, it may be possible to identify two different texts that share similar passages, which are paraphrases, not citations. The phenomenon of borrowing is very common in Buddhist texts. In order to identify and locate such cases, the large number of Tibetan texts at our disposal would need to be searched and compared. The scholarly implications promise to be far reaching, as this would enable the discovery of the history and emergence of texts and scriptures, and allow for the estimation of the popularity of certain texts that are cited more often (and to determine in which circles these are cited). Moreover, since in the case of the Tibetan canonical texts, translated material is being used, the translation and editorial practices can also be explored.

Before The Match:

a-g- da---ng gzhān dang, 'jig rten ph---a- rol srung ba dang, bcom l
ongs su srung b---a- dang- | -- nyan pa dang | mchöd pa -dang- | -- n
dan 'das kyi bstan pa srung ba dang, nyan-- tho-s kyi theg pa dang, -
yan tho-s kyi g-tam da---ng | rang- s-angs rgyas kyi gtam- -- dang |
ran---g sa---ngs--- rgyas kyi theg
theg pa chen po'i -gtam nyan---

The Match:

pa dang, - -theg pa chen po la yang dag -par zhugs pa'i gang zag ts
---pa dang | theg pa chen po la yang dag par zhugs pa'i gang zag ts

hul khriṃs dang ldn pa yon tan gyis -phyug pa,
hul -khriṃs dang ldn pa yon tan gyis phyug pa-

After The Match:

snod du gyur pa dang, ---- snod du ma gyur pa- nas mg-o- bregz te, ngu
grol ba dang | ri-gs pa'i spobs- pa can- dag la- mchöd pa -- da- -ng

r smri-g ---gi tshal bu gyon pa'i bar la- - srung bar byed, mchöd- - p
| de dag dang- l-han ci-g tu dga'- ba- dang -rtse ba- dang- y-ongs s

ar byed-- pa dang, mchöd- - rten-
u ' - -dri ba dang yongs su 'd

Fig. 2: Examples of quotations found. (a) The quotations with the 90th highest score, including its context, before and after. Line breaks have no significance. (Vertical bars serve as commas.) Two hues depict the two texts, red/pink for the Sūtrasamuccaya and blue/magenta for the Śikṣāmuccaya.

ngs su -rdzogs par bya'o snyam du brtag par bya'o,--- ,-----
ngs su rdzogs par bya'o snyam du brtag par bya'o || de bzhin du sb
-----khyim bdag gzhān yang byang chub sems dpa' rab tu -byung ba
yar te | khyim -bdag gzhān yang byang chub sems dpa' rab tu byung ba

dgon pa na gnas pas, bdag ci'i phyir dgon par 'ongs snyam -du brtag
dgon pa na -gnas pas- bdag ci'i phyir dgon par 'ongs snyam du brtag

par bya ste,--- -----'di ltar bdag ni 'jigs shing skrag -pa'i phyir
par bya -ste de yang 'di ltar bdag ni ---'jigs shing skrag pa'i phyir

Fig. 2: Examples of quotations found. (b) Part of the quotation with the fourth highest score, which is relatively clean except for some omissions.

a	b	c	d	e	g	h	i	j	k	m	n	o	p	q	s	t	u	w	y	z
23685	6	1	21	24	52	4	39	1	1	20	4	15	68	3	30	14	6	18	5	1
b	7679	8	11	2	32	9	1	0	16	5	12	73	6	17	0	0	0	0	0	0
c	0	8	2842	4	0	0	5	0	3	0	4	0	2	11	6	5	3	8	0	4
d	30	18	2	9371	2	42	12	4	2	5	15	10	55	3	12	16	35	10	8	0
e	25	1	0	4	4635	16	2	33	0	0	2	15	6	19	0	0	16	2	8	0
g	33	21	5	48	10	15099	4	19	0	8	23	13	27	13	3	15	42	5	12	0
h	8	4	6	10	4	9	4965	1	0	0	3	0	19	5	1	2	36	13	1	0
i	54	2	0	8	25	40	0	6084	0	0	4	7	15	23	0	30	31	2	25	0
j	1	0	1	0	1	3	0	446	0	0	4	3	0	0	1	3	0	0	0	0
k	1	11	2	2	0	6	0	0	0	1549	5	0	4	0	2	2	3	3	3	5
l	6	10	1	16	1	4	1	2	2	4	3038	6	14	0	10	12	9	5	0	0
m	18	5	0	8	5	25	1	0	0	3	8	4653	8	14	5	8	13	1	3	0
n	58	11	5	51	4	26	7	19	1	18	12	11613	3	49	6	38	7	6	0	16
o	59	0	0	2	17	9	2	15	0	0	1	1	6	5077	18	3	0	11	0	5
p	23	92	15	29	0	3	2	1	0	1	25	8	59	0	8473	3	4	2	1	0
q	4	4	2	12	1	10	3	0	1	11	8	3	2	17	0	7193	9	0	2	14
r	11	6	28	19	84	24	0	1	1	11	8	3	2	17	6	11	49	10151	0	0
s	23	0	0	15	1	16	3	2	0	2	12	1	7	0	3	11	3	2267	0	0
u	68	0	0	1	25	48	1	26	0	0	2	1	8	10	0	5	15	0	3936	0
v	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0
w	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
y	13	2	2	23	0	8	12	4	2	6	6	4	7	0	8	3	11	19	0	6
z	18	0	4	3	0	6	1	0	1	2	4	3	0	0	1	1	10	3	0	6719

Fig. 3: Substitution counts between the Sūtrasamuccaya text (rows) and the Śikṣāmuccaya (columns).

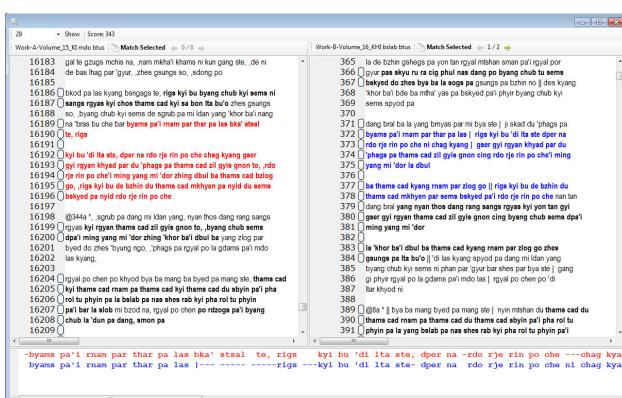


Fig. 4: A screenshot of the user interface, with the two matching texts displayed side by side. The text regions for which matching texts exist are emboldened. The first out of two texts that match the blue text are shown in red. This match has a score of 343 and is ranked 29th out of all matches. The panel at the bottom of the screen displays the two texts aligned character by character.

References

Barsky, M., Stege, U., Thomo, A., Upton, C. (2008): *A graph approach to the threshold all-against-all substring matching problem*. ACM Journal of Experimental Algorithms 12

Prasad, A.S., Rao, S. (2010): *Citation matching in Sanskrit corpora using local alignment*. In: Jha, G. (ed.): Sanskrit Computational Linguistics. Lecture Notes in Computer Science, Vol. 6465. Springer Berlin, Heidelberg 124–136

Supporting cross-media analyses by automatically linking multiple collections

Kleppe, Martijn

kleppe@eshcc.eur.nl
Erasmus University Rotterdam, Netherlands, The

Kemman , Max

Erasmus University Rotterdam, Netherlands, The

Introduction

Analysing media coverage across several types of media-outlets is a challenging task for Humanities researchers. Up until now, the focus has been on newspaper articles: being generally available in digital, computer-readable format, these can be studied relatively easily. Analyses of visual material like photos or television programs are however rarely undertaken. This poster presents the results of the PoliMedia project that aimed to showcase the potential of cross-media analysis by linking the digitised transcriptions of the debates at the Dutch Parliament with three media-outlets: 1) newspapers in its original format and lay-out of the historical newspaper archive at the National Library, 2) radio bulletins of the Dutch National Press Agency (ANP) and 3) newscasts and current affairs programmes from the Netherlands Institute for Sound and Vision (Kleppe et al., 2014; Kemman & Kleppe, 2013). The PoliMedia search user interface allows researchers to search through the debates and analyse the related media coverage via www.polimedia.nl. The main research question that can be addressed using PoliMedia is: *What choices do different media make in the coverage of people and topics while reporting on debates in the Dutch parliament since the first televised evening news in 1956 until 1995?* An advantage of PoliMedia is that the coverage in the media is incorporated in its original form, enabling analyses of both the mark-up of news articles as well as the photos in newspapers. PoliMedia demonstrates the application of Linked Open Data in the Digital Humanities: not only was a search interface developed for scholars, the data was published online and made publicly available via a SPARQL endpoint at data.polimedia.nl. This enables researchers to build customised tools that can support their specific research.

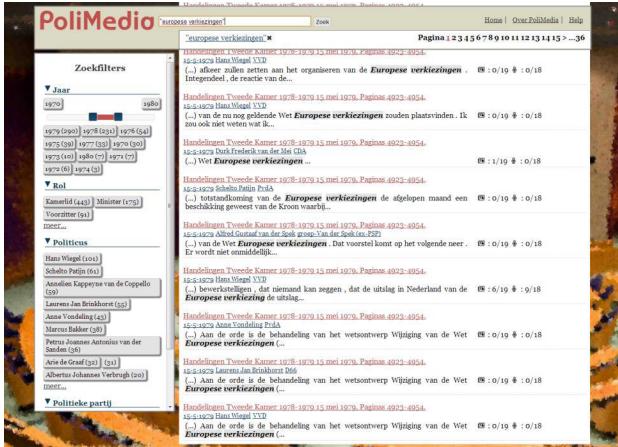


Fig. 1: Screenshot of the PoliMedia search results page

Method

The basis of PoliMedia lies in the minutes of the Dutch parliament from 1814-1995, containing circa 2.5 million pages of debates with speeches that have been OCR'd and thus allow full-text search. The minutes have been converted to structured data in XML form in previous research (Gielissen & Marx, 2009). For each speech (i.e. a fragment from a single speaker in a debate), we extract information to represent this speech; the speaker, the date, important terms (i.e. named entities) from its content and important terms from the description of the debate in which the speech is held. This information is then combined to create a query with which we search the archives of the newspapers, radio bulletins and television programmes. Media items that correspond to this query are retrieved, after which a link is created between the speech and the media item, using semantic web technologies (Juric, Hollink, & Houben, 2013). In order to navigate these links, a search user interface was developed, based on a requirements study with five scholars in history and political communication. During development, an initial version of this interface was evaluated in an eye tracking study with 24 scholars (Kemman, Kleppe, & Maarseveen, 2013).



Fig. 2: Screenshot of the PoliMedia debate page

Results

From an evaluation of a set of links to newspaper articles, it was found that the recall of the algorithm is approximately 62%, with a precision of 80% (Juric et al., 2013; Kleppe et al., 2014). However, no links to television programmes could be made. At this point we can make no conclusions about whether this was due to the size of the television dataset, the lack of full-text search or due to lack of suitability of the linking algorithm. Linking to television programs thus remains a question for future research. The combination of a search interface and a SPARQL endpoint resulted in PoliMedia

becoming the finalist of the Semantic Web Challenge 2013 en winning the first prize in the LinkedUp Veni Competition.¹ The poster presentation will be accompanied with a live demo of the system via www.polimedia.nl.

References

- Gielissen, T., & Marx, M. (2009). *Exemelification of parliamentary debates*. In *Proceedings of the 9th Dutch-Belgian Workshop on Information Retrieval (DIR 2009)* (pp. 19–25).
- Juric, D., Hollink, L., & Houben, G. (2013). *Discovering links between political debates and media*. In *The 13th International Conference on Web Engineering (ICWE'13)*. Aalborg, Denmark. doi:10.1007/978-3-642-39200-9_30 [<http://challenge.semanticweb.org/2013/finalists.html>](http://challenge.semanticweb.org/2013/finalists.html) & <http://blog.okfn.org/2013/09/17/linkedup-open-education-veni-competition-the-winners/>

Kemman, M., & Kleppe, M. (2013). *PoliMedia - Improving Analyses of Radio, TV & Newspaper Coverage of Political Debates*. In *Research and Advanced Technology for Digital Libraries* (pp. 401–404). Springer-Verlag Berlin Heidelberg. doi:10.1007/978-3-642-40501-3_46

Kemman, M., Kleppe, M., & Maarseveen, J. (2013). *Eye Tracking the Use of a Collapsible Facets Panel in a Search Interface*. In *Research and Advanced Technology for Digital Libraries* (pp. 405–408). Springer-Verlag Berlin Heidelberg. doi:10.1007/978-3-642-40501-3_47

Kleppe, M., Hollink, L., Kemman, M., Juric, D., Beunders, H., Blom, Oomen, J., Houben, G.-J. (2014). *PoliMedia - Analysing Media Coverage of Political Debates By Automatically Generated Links to Radio & Newspaper Items*. In M. D'Aquin, S. Dietze, H. Drachsler, M. Guy, & E. Herder (Eds.), *Proceedings of the LinkedUp Veni Competition on Linked and Open Data for Education*. Geneva, Switzerland: CEUR-WS.

LAUDATIO-Repository: Accessing a heterogeneous field of linguistic corpora with the help of an open access repository

Krause, Thomas
Humboldt-Universität zu Berlin

Lüdeling, Anke
Humboldt-Universität zu Berlin

Odebrecht, Carolin
Humboldt-Universität zu Berlin

Romary, Laurent
Humboldt-Universität zu Berlin, INRIA France

Schirmbacher, Peter
Humboldt-Universität zu Berlin

Zielke, Dennis
Humboldt-Universität zu Berlin

Open access to digital research data for historical linguistics enables a fruitful exchange of resources and methods. To achieve this goal the LAUDATIO-Repository provides long-term storage and open access to historical corpus linguistic data. We build a research data repository that on the one hand serves a set of well-defined scholarly communities' needs, but is on the other hand flexible enough to be used and extended to serve other communities not considered beforehand. In order to consider the scholar's needs, a clear understanding of the research data usage and research practice is required.

Digitalization and annotation of historical texts for linguistic purposes is a methodologically and technologically challenging task that consumes a lot of time and resources. Linguistic analysis uses various annotations and methods such as

multiple segmentations¹, normalizations², constituent syntax trees, dependency annotations and (semi-) automatic consistency checks³. Having the chance to either look up these methods and their results, the corpus itself, or to re-use existing corpora for further research may facilitate and enrich the research methods and possibilities of the linguistic community. An extensive documentation of the whole corpus preparation phase including annotation, tools and the resulting data allows a uniform and structured access to such a heterogeneous field of research data and thereby a first establishment of best practice standards.

There are many tools and formats that are used to annotate and process historical corpora, e.g. for token based annotations and parses. g.⁴,⁵). Thus the repository is open to all formats. The repository is based on the Open-Source Software Fedora 3.6⁶ and a custom web interface that uses its API. A single corpus can be stored and downloaded in multiple formats. A specific corpus version therefore comprises any number of data streams that capture the same dataset at a certain point in time. This mechanism is flexible enough to deal with annotation format diversity deriving from existing annotation tools such as EXMARaLDA, @nnotate or ELAN and as well as new formats (that have not been invented yet).

We have also developed a unified meta-model for (historical) text corpora that is used to describe heterogeneous corpus linguistic data and that can be modified according to future developments in the research field. Our repository will support all existing metadata versions for compatibility reasons. For each corpus in the repository a metadata documentation using this meta-model exists. Each corpus' metadata is automatically validated against the scheme while being imported into the repository. To enable a flexible metadata modeling we chose a customization of TEI XML with the help of an ODD⁷ specification. The meta-model indirectly refers to basic concepts of corpora.² Each corpus has a structured and uniform documentation which comprises information about the corpus creation, the documents and every kind of annotation. For instance the authorship of each annotation layer is reported for citation and copyright reasons. Furthermore the whole corpus preparation process is covered as well.

In order to support re-use of existing corpora and to reduce duplicate work, researchers can search, download and import new corpora. Additionally, new annotations or documents can be added as new versions of existing corpora. The detailed metadata documentation helps researchers to understand the corpus and its context. Moreover LAUDATIO provides a faceted and free-text search for certain annotation methods, tools or annotation values (see⁸ for an extensive discussion of faceted search). ElasticSearch³ is used as a technical backend.

The flexible technical infrastructure and the extensive corpus documentation of LAUDATIO will be presented in the demo session.

Footnotes

¹ Multiple segmentations or tokenizations are used to represent different, possible conflicting, interpretations of the smallest units in a corpus.

² The current documentation of the metadata can be accessed under korpling.german.hu-berlin.de/schemata/laudatio/doc/S6/corpus

³ ElasticSearch website, www.elasticsearch.org, accessed 29 Oct. 2013

References

1. Krause, Thomas, Lüdeling, Anke, Odebrecht, Carolin, Zeldes, Amir (2012) *Multiple Tokenization in a Diachronic Corpus. Exploring Ancient Languages through Corpora Conference (EALC)*, 14.-16.Juni 2012.
2. Bollmann, Marcel, Petran, Florian, Dipper, Stefanie (2011) *Rule-Based Normalization of Historical Texts*. Proceedings of Language Technologies for Digital Humanities and Cultural Heritage Workshop Hissar, Bulgaria, 16 September 2011. pp. 34–42.
3. Dickinson, Markus, and W. Detmar Meurers. . *Detecting errors in part-of-speech annotation*. Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-VOLUME 1. Association for Computational Linguistics.
4. Schmidt, Thomas (2009) *Creating and Working with Spoken Language Corpora in EXMARaLDA*. In: Lyding, V. (ed.) LULCL II: Lesser Used Languages & Computer Linguistics II. pp. 151-164.
5. Brugman, Hennie, Russel, Albert (2004). *Annotating Multimedia/ Multi-modal resources with ELAN*. In: Proceedings of LREC 2004, Fourth International Conference on Language Resources and Evaluation.
6. Lagoze, Carl, et al. (2006) *Fedora: an architecture for complex objects and their relationships*. International Journal on Digital Libraries 6.2: 124-138.
7. Burnard, Lou, Rahtz, Sebastian (2004) *RelaxNG with Son of ODD*. Extreme Markup Languages Proceedings 2004. Montréal, Québec.
8. Tunkelang, Daniel (2009) *Faceted search*. Synthesis Lectures on Information Concepts, Retrieval, and Services 1.1. pp. 1-80.

Detecting Linguistic Signal in Cather's Early Journalism: Polishing the Bibliography

Kumari, Ashanka

akkumari@crimson.ua.edu
University of Nebraska-Lincoln

Lawton, Courtney

coco.lawton@gmail.com
University of Nebraska-Lincoln

McCue, Carmen

carmen.mccue@msn.com
University of Nebraska-Lincoln

Moreno, Jose Luis

josebakmoreno@gmail.com
University of Nebraska-Lincoln

Thomas, Grace

gracethomas.unl@gmail.com
University of Nebraska-Lincoln

Background and History

Willa Cather, a novelist whose works include *My Antonia* and *O Pioneers!*, served her literary apprenticeship writing drama, music, and book reviews for literary magazines and small-town newspapers between 1893 and 1903. Most of these articles were unsigned or pseudonymous, but were attributed to Cather usually based on circumstance and opportunity, reference, and style – something William Curtain failed to define beyond “the style and swing [being] unmistakably Cather’s” (Curtain 972). Authorship attribution is hardly a new field for digital humanities. David Hoover, Matthew Jockers, Hugh Craig, Patrick Juola and others have done work on literary authorship attribution in which they examined and refined the different approaches (e.g. function words and lexical entropy) to the authorial question. Our project focuses on a small corpus of short texts, reflecting our attempt to verify authorship attribution in Willa Cather’s early unsigned, attributed journalism from 1893 to 1903.¹

Early efforts to create a bibliography of all of Cather’s writing, including documentation of her journalism, were complicated by Cather’s pseudonymous and anonymous publications. In 1950, John Hinz created an initial list of Cather’s pseudonyms, including Mary K. Hawley, Clara Wood Shipman, and Helen Delay, her most famous pseudonym (Hinz 201). In 1966,

Bernice Slote published the first definitive bibliography of Cather's early nonfiction writings in *The Kingdom of Art* where she uncovered 44 previously unattributed columns she believed Cather wrote for the *Nebraska State Journal* between 1895 and 1896. In 1970, William Curtain compiled *The World and the Parish*, a two-volume bibliography of Cather's journalistic writings. Joanna Lathrop followed up with a 1975 checklist. Joan Crane's 1982 bibliography of Cather's work is a departure point for much recent scholarship, including the online *Willa Cather Archive* (cather.unl.edu).

Tim Bintrim's 2004 dissertation attributed 19 additional pseudonymous articles to Cather during her Pittsburgh years and successfully challenged at least two pseudonyms proposed by Hinz. In 2013, Kari Ronning of the University of Nebraska-Lincoln and Robert Thacker of St. Lawrence University cleared up confusion about a purported Cather pseudonym "Clara Wood Shipman." It is not a Cather pseudonym, something Slote had hinted about in a footnote in *The Kingdom of Art* (Slote 28).

In this research, we investigate Hinz's original attributions as well as those suggested by scholars that came after him. We apply tools and techniques of authorship attribution and stylometrics in order to assess the extent to which unsigned works and works attributed to Cather possess a stylistic and linguistic voice that is consistent with her known works.

Methodology

We approached the initial problem of authorship attribution first by breaking down the content of each of Cather's journal articles to the word level. Our sample consisted of 158 texts. Of these, 126 texts were made available through the *Willa Cather Archive*. The remaining 32 texts were articles written by Fanny Fern (Sarah Parker Willis), which were used as a control group in order to establish a clear stylistic signal that was not Cather's. We chose Fern's journalistic work not only because it was available in digital form, but also because Fern's work constitutes a body of work that is comparable to Cather's in terms of genre, size, number of works, and available metadata. At the same time, the Fern material is distinctive enough chronologically and geographically to avoid any artistic overlap.

Our total corpora consisted of 216,506 words, or tokens, 192,062 belonging to the Cather corpus and 24,444 belonging to the Fern corpus. The identification of unique token types resulted in 20,868 types: 15,656 appearing only in Cather's corpus, 5,212 appearing only in Fern's corpus, and 3,591 shared between the two. Our algorithm then employed simple relative frequency across the corpora in order to account for any discrepancies in article length. The resulting plots are based on the occurrence of the most frequent words in the corpora, such as "the", "a", and "of," above an arbitrary threshold. Instead of imposing structural rules, we practiced unsupervised clustering by allowing the algorithm to group the works itself based on the frequency of these common words. This created hierarchically clustered dendrogram plots, which serve to provide a visual analysis. At first, we seemed to detect a clear linguistic signal for Cather. When we introduced the control set, Fern's works also tended to cluster together in our initial dendograms (see figure 1).

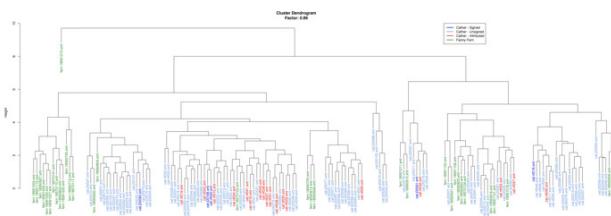


Fig. 1: Example of one of the original, first-run results dendrograms (here, run with a threshold factor of 0.89).

After closely reading the pieces, we discovered that there was a substantial amount of quoted material from other authors (such as William Shakespeare and Rudyard Kipling), which might have skewed the accuracy of our approach. In order to

produce the most reliable results, we manually stripped the quoted material from the files.

Observations

After carefully examining the dendograms from our analysis, we have evidence that the works analyzed were, correctly attributed to Cather, but we cannot be sure to have detected her style or signal.² Additionally, we found the third-party quotations in Cather's work had no significant effect on the clustering of works. Contrary to our expectations and conventional wisdom, the same works consistently clustered before and after the extraction of quoted material. No matter which frequency threshold we used to produce the hierarchically clustered models, the results (see figures 2 and 3) did not offer a solid pattern for determining Cather's early journalistic style or identify possible misattributions.

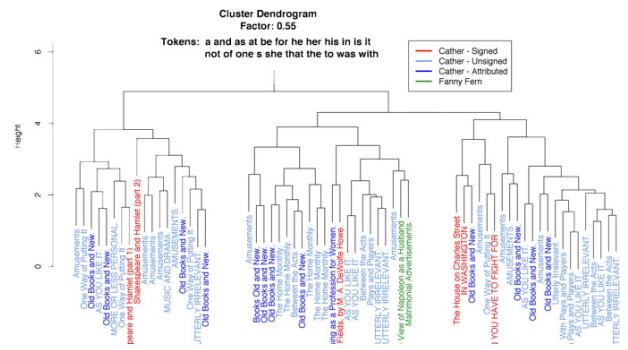


Fig. 2: One of the final dendograms (here, run with a threshold factor of 0.55).

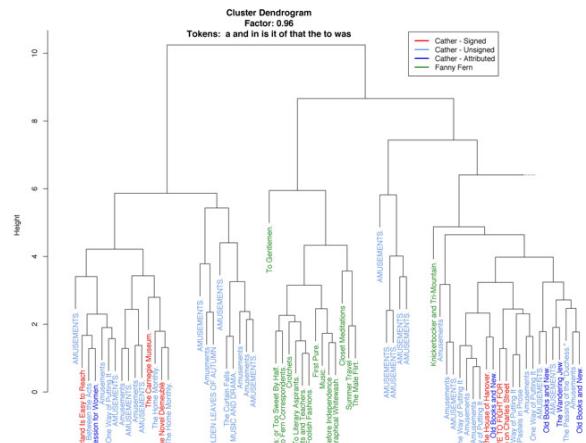


Fig. 3: One of the final dendrograms (here, run with a threshold factor of 0.96).

However, we can claim the veracity of the previously attributed works achieved through traditional scholarship by employing this computational method. With little controversy, the clusters formed by signed or consistently attributed pieces support such articles as likely to have been written by Willa Cather. Our control set consisting of Fanny Fern's work generally tended to cluster together as a set, indicating that our analysis distinguished between the two authors with high reliability.

Conclusions and Further Research

Our initial efforts at detecting a Cather signal to dispute the attribution of anonymous and pseudonymous texts were not successful. While we can be definitive enough to support Curtain, Bintrim, Slote, and Crane's attributions, our results cannot be considered reliable enough to dispute any dubious attributions. Burrows, Jockers, Hoover, and Craig suggest

machine reading of texts and unsupervised clustering of shorter texts and smaller corpora is still better than working by mere chance. The question is, how much better?

The context of our work seemed to be appropriate for quantitative methods. We set out to allow a machine to do that for which few people have little patience: counting the frequency of function words. As Hoover suggested, "Only when external evidence fails is it reasonable to apply quantitative methods, and the presence or absence of a closed set of possible authors and differences in the size and number of documents available for analysis are usually more significant than the kind of text involved." The external evidence seemed compelling: Cather's output as a young journalist was copious, and scholars using traditional methods were able to identify more than three pseudonyms incorrectly attributed to her. We analyzed a closed set of authors, Willa Cather and Fanny Fern, so we expected to find more attribution errors, based upon the timbre of early Cather scholarship, which often smacked of hagiography. What we were lacking was a larger corpus.³ This project indicates the need to digitize more of Cather's early work to further clarify the historic bibliography using computational stylometry.

Bibliography

- **Bintrim, Timothy W.** (2004) *Recovering the Extra-Literary: The Pittsburgh Writings of Willa Cather*. Diss. Duquesne University. Print.
- **Burrows, J.F.** (2004). "Textual Analysis." *A Companion to Digital Humanities*. Ed. Susan Schreibman and Ray Siemens. Oxford: Blackwell, 2004. Web.
- **Craig, Hugh** (2004). "Stylistic Analysis and Authorship Studies." *A Companion to Digital Literacy Studies*. Ed. Susan Schreibman and Ray Siemens. Oxford: Blackwell. Web.
- **Crane, Joan** (1982). *Willa Cather: A Bibliography*. Lincoln, NE: U of Nebraska P. Print.
- **Curtain, William**(1893-1902). *The World and the Parish: Willa Cather's Articles and Reviews* . Lincoln, NE: U of Nebraska P, 1970. Print.
- *Fanny Fern in the New York Ledger*(2013). Center for Digital Research in the Humanities at the University of Nebraska-Lincoln. Web. 9 Dec. 2013.
- **Hinz, John P** (1950). "Willa Cather in Pittsburgh." *The New Colophon*. New York: Duschnes Crawford Inc. Print.
- **Hoover, David L** (2004). "Quantitative Analysis and Literary Studies." *A Companion to Digital Literary Studies*. Ed. Susan Schreibman and Ray Siemens. Oxford: Blackwell. Web.
- **Jewell, Andrew, and Janis Stout** (2013). *The Selected Letters of Willa Cather*. New York: Knopf. Print.
- **Jewell, Andrew**, ed. *The Willa Cather Archive*. U. of Nebraska-Lincoln, (2004-2013). Web. 9 Dec. 2013. Web.
- **Jockers, Matthew L. and Daniela M. Witten**(2010). "A Comparative Study of Machine Learning Methods for Authorship Attribution." *Literary and Linguistic Computing*, 25.2 : 215-224. Web. 9 Dec. 2013.
- **Jockers, Matthew L., Daniela M. Witten, and Craig S. Criddle**(2008). "Reassessing Authorship of the Book of Mormon Using Delta and Nearest Shrunken Centroid Classification." *Literary and Linguistic Computing*, 23.4 : 465-492. Web. 9 Dec. 2013.
- **Jockers, Matthew L.** (2013) "Testing Authorship in the Personal Writings of Joseph Smith Using NSC Classification." *Literary and Linguistic Computing*. 28.3: 371-381. Web. 9 Dec. 2013.
- **Jockers, Matthew**. (2013). *Text Analysis with R for Students of Literature*. 2 Sept. TS.
- **Juola, Patrick**. "Authorship Attribution." *Foundations and Trends in Information Retrieval* 1.3 (2006): 233-334. Web. 1 Dec. 2013.
- **Lathrop, JoAnna** (1975). *Willa Cather: A Checklist of her Published Writing*. Lincoln, NE: U of Nebraska P. Print.
- **Ronning, Kari**. "Re: theatres in Lincoln." Message to the author. 9 Dec. 2013. E-mail.
- **Slote, Bernice**. *The Kingdom of Art*. Lincoln, NE: U of Nebraska P, (1966). Print.
- **Van Dalen-Oskam, Karina** (2013). "Epistolary Voices: The Case of Elisabeth Wolff and Agatha Dekken." *Digital*

Humanities 2013 Lincoln NE. Eds. Katherine Walker and Kenneth Price. Lincoln, NE: Center for Digital Research in the Humanities. Web. 9 Dec. 2013.

References

1. This research began in a class taught by Matthew Jockers and has continued under his direction as a project of the newly formed Nebraska Literary Lab at the University of Nebraska-Lincoln.
2. If we had found Cather's signal, it would be best described as the distribution of works that clustered together, whereby some elements of the clusters that have the same frequency consequently have the same style.
3. According to Andrew Jewell, editor of the online Cather Archive, there are a total of approximately 600 pieces of journalism written by Cather. The Cather Archive has digital scans of about 400 of those. About 230 are transcribed in some fashion, and about 206 are in some kind of XML format.

Annotating texts with ontologies, from geography to persons and events

Lana, Maurizio

m.lana@lett.unipmn.it
Università del Piemonte Orientale

Ciotti, Fabio

fabio.ciotti@uniroma2.it
Università di Roma Tor Vergata

Magro, Diego

magro@di.unito.it
Università di Torino

Peroni, Silvio

essepuntato@cs.unibo.it
Università di Bologna

Tomasi, Francesca

francesca.tomasi@unibo.it
Università di Bologna

Vitali, Fabio

fabio@cs.unibo.it
Università di Bologna

1. Introduction

1.1. Overview

Geolat - geography for latin literature aim is to annotate every placename in the latin literature using a geographical ontology for the ancient world which must be built from scratch. This will allow scholars but also students and citizens to start reading texts choosing a specific area or place they are interested to. The starting point is the Latin literature because of its founding meaning for the European culture ¹.

But this same model can be applied to every other literature because nothing in it is language-dependent. Should a scholar be able to browse a map of Europe (or of the world) to choose a specific place and start examining which authors in which works wrote about it is a completely new way or conceiving the study of literature.

But this model can be expanded from place names to persons and events, allowing to browse texts by the names of people they contain or by types and subtypes of events. That is texts are meant as collections of information. Not always and not necessarily factual information because literary texts can speak of mental, fictional representation of places or people.

Nevertheless what texts say often constitutes the canvas of what real places and people and events actually are or were.

Concurrent 'sayings' about the same place, person, or event, can be managed strictly connecting the annotation to the textual object containing the statement(s). So an ontology of textual objects is needed.

1.2. Methodology

The main methodology is that of starting from a prototype where a geographical ontology is used to annotate place names in latin texts. At this level a startup funding is sufficient to build a prototype showing an example of what could be obtained.

Then the prototype can be expanded:

- adding more literatures;
- and/or adding more ontologies.

In both cases a great effort is needed to accomplish the task and adequate funding is indispensable.

As the ontology is the core of the project something must be said about it. Conceptually, the annotation mechanism is based on EARMARK². EARMARK is an OWL 2 DL ontology that defines document meta-markup (elements, attributes, comments and text nodes), and it can be used to express facts about the inherent semantics of the markup elements and of the content of a text³ according to a precise semiotic model⁴. Thus, using EARMARK it is possible to create annotations on (plain or marked up) documents and document fragments according to a multilayer architecture (that can be aligned, in principle, with the Open Annotation Data Model⁵) and allows one not only to specify by an annotation the referent of a geographic place name, but also to provide a set of data describing such referent. All of the entities involved in an annotation are identified by an IRI and both the annotation, its metadata and its content are specified according to the RDF data model. In this way, all these pieces of information can be published on the Web of Data following the Linked Data principles⁶. The semantics of these data are explicitly expressed by means of formal ontologies, thus allowing the Geolat system itself as well as external applications to capture (at least partially) the data, to draw inferences on them and facilitating data integration. Given the specificity of the domain and of the task, a Geolat ontology has been built: We will discuss it and its relationship with other geographical ontologies widely used in the Linked Data world, such as Geonames⁷.

2. Getting Started

What is here described is a running project - geolat - with some possible extensions.

2.1 The Geolat Project

Geolat project is intended to build a complete digital library of classical and late latin texts, to annotate them in every place name and then to offer a double interface to browse the texts: a cartographic one, and a mixed one (textual + cartographic). The first one allow users to choose a region on a map and to obtain a list of authors, and works, and place names pertaining to that region; the second one allows to search for a specific place whose name is written in a text field; the search will show a list of passages, a display of places onto a map, histograms of presence of place names in various sections of the related works, list of other cooccurrent place names.

2.2 The Geolat Prototype

A prototype of geolat system will be built for the first months of year 2014 and will be available online at the address <http://www.geolat.eu>. It will offer a small digital library with some ten latin texts. The placenames contained in these texts will be manually annotated using a geographical ontology specifically conceived for the classical (latin) world and literature. A

cartographic interface will allow to browse the texts clicking regions or places showed onto a map.

2.3 Building the Digital Library

As it was said, the first step is having a digital library of texts. The texts in the library will be annotated using a mixed approach: the in-line annotation is based on the TEI standard, and links to an external ontology which describes geographical entities how the ancient Romans did: a forest is the home of a nymph, a river is sacred to a goddess, a city can have more than one founder, and so on.

2.5 Browsing the Texts through a Map-based Interface

The map-based interface is the more complex part of this project because of its novelty and because of some technical problems which must be solved (e.g. finding geographical places in a given radius from another one).

3. Going further

Project geolat can go further in two non mutually exclusive ways:

- adding literatures
- adding ontologies (that is, types of entities).

3.1 More Literatures

Adding more literatures to the initial latin one would allow to build distinct layers of literary interpretation and to extend the scope of analysis and queries. For instance we could ask in which texts (and inside them in which passages) of Latin and French literature we find contextual references to the city of Appida and to the person of Caesar.

From the technical point of view different literatures mean different digital libraries which must define a formal protocol for interoperability. There are some basic conditions to assure this interoperability (use of UTF-8 characters encoding; adoption of at least a basic markup - preferably in XML - describing and identifying its structure: title, sections, subsections, but also paragraphs, sentences, words; adoption of open software systems).

But the only way to assure the level of interoperability we envisage to heterogeneous collections is agian based on sharing conceptual and ontological assumption.

3.2 More Ontologies

Adding more ontologies, that is more types of categories, would offer a more complex access to the texts and the knowledge they contain.

3.2.1 Persons

Recognizing and semantically annotating geographical entities in only a first step to the goal of providing an innovative access to literary works. As witnessed by the work of Francesca Tomasi⁸, persons and characters play a central role in many literary productions. In particular, we will describe how the system support the identification and the formal representation of semantic relationships between texts and persons (or characters), among persons themselves and, in general, between persons and other resources and how the EAC-CPF Ontology⁹ can be exploited to this aim. The analysis of the context, as conceived in the archival domain, is a fundamental approach to connect people to documents on the basis of the role or function covered. Roles are the key tool to manage self-explanatory relationships. For this reason the Pro Ontology

¹⁰ will be also considered. We will also describe how the integration of the geographical and the personal perspectives in a homogeneous framework enhances the benefits provided by both.

3.2.3 Events

Since people participate to events and many relevant events take place in the real world (meetings, battles, travels, etc.), we believe that the notion of event is a sort of conceptual glue between geographical places and people that should be captured within our system. We will discuss this issue.

Events are of different types and are bound to time, to a specific point or duration in time.

Good examples of ontologies for events are the "Simple Event Model":

lov.okfn.org/dataset/lov/details/vocabulary_sem.html¹¹ or the "Linking Open Descriptions of Events": linkedevents.org/ontology

3.2.4 Textual Objects

Moreover, the work of Peroni and Vitali on semantic publishing^{12 13} proved how the semantic technologies can be exploited in order to enrich the meaning of documents published on the Web by specifying meaningful relationships between them¹⁴. We will discuss how this approach can be applied in to the digital library provided within the discussed system, in order to complement it with a rich network of semantically linked resources connecting both the productions by themselves and their contents (places, persons, characters and events) in a coherent semantic graph.

References

1. **M. Lana, Geolat:** *Geography for Latin Literature*, in (forthcoming) ISAW papers 7, Current Practice in Linked Open Data for the Ancient World Editors: Thomas Elliott, Sebastian Heath, John Muccigrosso, sfsheath.github.io/lawdi-publication/isaw-papers-7.xhtml
2. **Di Iorio, A., Peroni, S., Vitali, F.** (2011). *Using Semantic Web technologies for analysis and validation of structural markup*. In International Journal of Web Engineering and Technologies, 6 (4): 375-398. Olney, Buckinghamshire, UK: Inderscience Publisher. DOI: 10.1504/IJWET.2011.043439
3. **Peroni, S., Gangemi, A., & Vitali, F.** (2011). *Dealing with markup semantics*. In Proceedings the 7th International Conference on Semantic Systems (I-SEMANTICS 2011): 111–118. New York, New York, US: ACM Press. DOI: 10.1145/2063518.2063533
4. **Picca, D., Gliozzo, A. M., & Gangemi, A. (2008).** *LMM: an OWL-DL MetaModel to Represent Heterogeneous Lexical Knowledge*. In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 08). Marrakech, Morocco: European Language Resources Association (ELRA). ISBN: 2-9517408-4-0
5. **Open Annotation Data Model**, W3C Open Annotation Community Group 2013, www.openannotation.org/spec/core
6. **Tom Heath and Christian Bizer**, *Linked Data: Evolving the Web into a Global Data Space*. Synthesis Lectures on the Semantic Web: Theory and Technology, Morgan & Claypool 2011
7. **Geonames Ontology**, www.geonames.org/ontology/documentation.html
8. **J Tomasi, Francesca**. (2013). *Digital editions as a new model of conceptual authority data*, JLIS.it 4.2 21-44
9. **J Mazzini, Silvia, and Francesca Ricci**. (2011). *EAC-CPF Ontology and Linked Archival Data*. In Semantic Digital Archives (SDA) Proceedings of the 1st International Workshop on Semantic Digital Archives. ceur-ws.org/Vol-801/
10. **Shotton David, Peroni Silvio**. (2010). *Pro - The Publishing Roles Ontology* purl.org/spar/pro/ (last modified 2013-05-15)
11. **Willem Robert van Hage, Véronique Malaisé, Roxane Segers, Laura Hollink, Guus Schreiber**. *Design and use of the Simple Event Model (SEM)- Web Semantics: Science, Services and Agents on the World Wide Web*, Volume 9, Issue 2, July 2011, Pages 128–136 dx.doi.org/10.1016/j.websem.2011.03.003 Postprint at: www.cs.vu.nl/~guus/papers/Hage11b.pdf
12. **Peroni, S., Shotton, D., Vitali, F.** (2012). *Scholarly publishing and the Linked Data: describing roles, statuses, temporal and contextual extents*. In Presutti, V., Pinto, H. S. (Eds.), Proceedings of the 8th International Conference on Semantic Systems (i-Semantics 2012): 9-16. DOI: 10.1145/2362499.2362502
13. **Peroni, S., Shotton, D., Vitali, F.** (2012). *Scholarly publishing and the Linked Data: describing roles, statuses, temporal and contextual extents*. In Presutti, V., Pinto, H. S. (Eds.), Proceedings of the 8th International Conference on Semantic Systems (i-Semantics 2012): 9-16. DOI: 10.1145/2362499.2362502
14. *Semantic Publishing and Referencing Ontologies*: purl.org/spar

Enduring Traces: Exploring correspondence from the archives of Canadian modernism using digital tools and methods

Lang, Anouk

anouk.lang@gmail.com

University of Strathclyde

This project uses Neatline and Gephi to demonstrate how digital visualization tools can bring to light new dimensions of modernist studies and periodical studies. Drawing on metadata from as-yet-undigitized letters between various Canadian writers and editors, the project uses geospatial visualizations and network diagrams to interrogate the literary networks and geographical patterning of authors associated with the little magazine *Contemporary Verse*. In addition to the printed poster, I will have a laptop running Neatline and Gephi, which will allow attendees to interact with the data by exploring it through maps, timelines and network diagrams.

Contemporary Verseran from 1941-1953 and was one of the only vehicles for the publication of modern poetry in mid-century Canada. It was edited by Alan Crawley (1887-1975), whose voluminous correspondence – lodged in various archives across Canada – is an enormously rich resource for the study of the social networks through which poetic currents and aesthetic influences developed. To read it is to get a clear sense of the importance of Crawley in brokering relationships between writers and publishers, shaping the poems that contributors submitted to the journal, and encouraging young writers – particularly younger women who faced considerable difficulties breaking into male-dominated networks for publication and critique – to see their poetry as something worth pursuing.¹

The letters tell an important story of pre-digital cultural empowerment that digital humanities approaches are particularly well suited to uncovering, given that the volume of correspondence lends itself to the kind of distant reading that is made possible by computational analysis of prosopographical and geographical metadata. The new perspectives opened up on this archival material by digital methods are also timely, as scholars of Canadian literature are increasingly engaged in challenging canonical accounts of modernism's development within the country,² something which has coincided with a postcolonial turn of sorts within digital humanities more generally.³

The poster, which reports on work in progress into the Crawley letters, is focused around two research questions:

1) How does geography inflect the development of modern poetry in Canada?

Each letter was entered as a record in Omeka and then visualized geospatially using Neatline (www.modmaps.net/mm/neatline/show/crawley-letters). Neatline was configured so that as the timeline slider was moved from earlier to later dates, the map display showed the growth and geographical distribution of Crawley's network of correspondents. Displaying the correspondence in this way enables investigation of spatial questions that arise at a range of scales from the national to the local, for example the extent to which Crawley's location on the west coast was able to challenge the dominance of literary networks in the eastern cities of Montreal and Toronto, and the effect of Vancouver's geography on his ability to participate in literary activities.

2) What is the relationship between literary networks and cultural production?

Gephi was used to create a directed graph with 25 nodes, representing Crawley, various authors with whom he corresponded, other journal editors, and the women who did the bulk of the administrative work for *Contemporary Verse*. As with the Neatline map, the network diagrams generated by Gephi do not give a complete picture of Crawley's network of correspondents as archival work is still ongoing, but nonetheless some preliminary suggestive patterns emerge. Crawley's initial correspondents are more likely to be women, for instance, but as the journal accrues prestige, more established poets become interested in submitting to it, and these poets were more likely to be men. Such a narrative is clearly open to critique – for example it raises methodological questions about which letters were included – but this is a valuable process for raising questions about the partiality of the "data" on which existing narratives about the development of Canadian modernism have relied.

This project forms part of **EMiC: Editing Modernism in Canada** (editingmodernism.ca). It is also being developed in association with **TCLL: Twentieth Century Literary Letters** (www.modmaps.net/tcllp), a collaborative project in its early stages which aims to build a digital infrastructure for the discovery, analysis, and visualization of the metadata from a wide range of epistolary materials relating to twentieth century literary figures. As a founding member of TCLL, I am keen to find other scholars working on letter collections who would be interested in joining the project, and one of the aims of showcasing this work at DH2014 is to make connections with others whose metadata could be productively brought into conversation with existing TCLL materials.

References

1. **McCullagh, Joan.** *Alan Crawley and Contemporary Verse*. Vancouver: U British Columbia P, 1976. Print; Robertson, George. "Alan Crawley and Contemporary Verse." *Canadian Literature* 41 (1969): 87–96. Print; Wilson, Ethel. "Of Alan Crawley." *Canadian Literature* 19 (1964): 33–42. Print.
2. **Irvine, Dean** (2011). *Spectres of Modernism*. *Canadian Literature* 209: 6–10. Print.
3. **Koh, Adeline, and Roopika Risam** (2013). *Open Thread: The Digital Humanities as a Historical "Refuge" from Race/Class/Gender/Sexuality/Disability?* Postcolonial Digital Humanities. 10 May 2013. Web. 1 Nov. 2013.

Cultivating the Public Philosophy Journal

Long, Christopher

The Pennsylvania State University, United States of America

Fisher, Mark

The Pennsylvania State University, United States of America

Rehberger, Dean

Michigan State University

Abstract

In this poster session, we present the project of the Public Philosophy Journal and our plans for cultivating a community of engaged scholars to sustain it. The presentation will explain our motivations for designing the journal to perform public philosophy as its mode of publication, highlight the journal's role as a hub for community-sourced curation and review of existing work, and introduce our model for the collaborative writing and editing of publicly engaged scholarship. We will draw attention to common aims of differing conceptions of public philosophy, and discuss how the PPJ will leverage digital media in promoting both reasoned deliberation concerning the public good and the modeling of virtues of thought, expression, and action within the public sphere.

Introduction

The last two decades have seen an increase in the urgency with which public universities and professional societies in the United States have turned to reflect on the culture of isolation and detachment that has come to characterize much of contemporary academic work. Nowhere, perhaps, has this reflection been more necessary in making the case for continued institutional and public support than in the discipline of philosophy. While changes in the academic culture of the discipline have been significant over the last twenty years, the model of publication that continues to dominate there serves to reinforce, rather than to undermine, the misperception that philosophical activity is essentially divorced from the concerns of public life. This model of publication remains valuable insofar as it rewards carefully researched and thoughtful work that meets the high intellectual standards of respected members of the discipline. However, the process through which work is prepared and approved for print publication is time-consuming, expensive, and constrained by a review process that favors technical work directed at problems framed by the discipline itself over work that takes up the shared responsibilities of public citizens.

Ultimately, existing modes of scholarly publication are not well suited to provide timely, open, public access to philosophically informed work that addresses widely shared public concerns. They are similarly constrained when it comes to providing the general public with examples of the discursive and collaborative processes through which philosophically rigorous work is produced.

The emergence of digital communication has brought important changes to the way philosophers produce scholarly work. Traditional publishers are also recognizing the opportunities digital media present for making scholarly work more widely available to various publics. Drawing on these developments, the Public Philosophy Journal (PPJ) is designed to leverage new modes of digital scholarly communication to produce and publish work that is equally responsive to the demands of public discourse and to the rigorous standards of academic publication.

Community-Sourced Scholarship

'Community-sourcing' stands as an alternative to the increasingly popular practice of 'crowd-sourcing'. Rather than leveraging the resources of a loosely connected crowd located somewhere 'out there', the PPJ seeks to draw upon resources that already exist within groups of integrated, overlapping communities. These communities are made up of practitioners who are oriented towards common public concerns. In his book *The Public and Its Problems*, Dewey captures the public spirit

that we see animating the “community-sourcing” approach of the PPJ:

To learn to be human is to develop through the give-and-take of communication an effective sense of being an individually distinctive member of a community; one who understands and appreciates its beliefs, desires and methods, and who contributes to a further conversion of organic power into human resources and values.

As a community in which one learns to be human in this sense, the PPJ’s publication practices will seek to curate and amplify the efforts of community members, thereby generating innovative connections across already existing communities, and carrying on the traditions of dialogue and rigorous scholarship within the forums emerging in the digital age.

Curating digital conversations from around the web through existing web-crawling technologies, a network of graduate students in philosophy and a wider community of engaged citizens, the PPJ will identify contributions that might be further developed for scholarly publication in the journal. Authors of these contributions will then be invited into a collaborative, developmental writing space in order to prepare their work for open peer review. This open peer review process will include a system for reviewing and credentialing reviewers and incentives for careful reading and for consistent and thoughtful commenting. Accepted articles will be openly published in the journal together with invited responses to the reviewed work and links to the sources from which the articles developed. These practices of community-sourced scholarship are designed to enrich both public discourse and philosophical scholarship.

Public Philosophy

This project is situated within a growing network of people who identify themselves as doing a kind of public philosophy. In *Practicing Public Philosophy*, Sharon M. Meagher and Ellen K. Feder report on three general positions on public philosophy taken by practitioners. The first identifies philosophical practice as a public good that should be practiced in the public sphere. The second frames public philosophy as philosophy aimed explicitly at benefitting the public. The third focuses on philosophical analysis as having a liberatory value in relation to existing structures of power. We agree with the authors that these three positions are not mutually exclusive. We also go beyond this claim in suggesting that the new era of digital media and communications provides us not only with a set of useful tools for enabling work that is described in these ways, but also with a set of opportunities and responsibilities with respect to the unifying theme that relates these positions to one another within a common project.

We are focusing on a public sphere that exists not only ‘beyond the ivory tower’ but also beyond those physical spaces of public assembly that have long served as its rhetorical contrasts (e.g., cafes, bars, and theaters, to name only a few). We are also encouraging a view of our own responsibility to act as collaborative user-designers of the technological infrastructure required to generate and sustain this sphere in ways that are productive of the goods we seek in practicing philosophy. The online, open access, open peer review structure of the PPJ will enable it to model the public practice of philosophy, in a way that is explicitly aimed to benefit the various public spheres in which its members participate, and that serves to liberate practitioners from at least some of the existing constraints on its acceptable outcomes. The design of the journal itself will be guided by the idea that the most urgent issues of public concern are best addressed in and through public scholarly communication that puts the virtues of responsible public discourse into practice.

Public Deliberation

We share with the community of scholars and practitioners that has emerged around the general topic of ‘online deliberation’ the hope that emerging digital media can be

leveraged to promote and sustain reasoned, purposeful, and interactive communications surrounding issues of public urgency. We also share the concern that other commercial and academic interests may seek to prevent these potentially transformative media from being used to their fullest potential. Recent scholarship on communication and deliberative discourse in the digital age provides us with both a plurality of models for facilitating productive public deliberation and some emerging consensus around general best practices for avoiding the kinds of digression that are so common both in online and in more traditional communication contexts.

The self-understanding of the community, cultivated through open shared public dialogue, is best positioned to shape the norms and guidelines according to which the journal puts public philosophy into practice. Accordingly, it is as important to us to community-source decisions about the policies to which contributors to the journal will be subject, and concerning the qualitative standards against which reviewers will themselves be reviewed, as it is to draw on the subject-matter expertise of community members in making decisions concerning the quality of the contribution made by any particular article under review.

This opens us up to the criticism that we are being overly optimistic concerning the community’s capacity for fair, equitable, and open self-legislation. However, we would prefer to admit that possibility while working to provide examples of excellent public scholarly deliberative practices, than to prejudge the case in a way that simultaneously underestimates the community and undermines the public spirit of the project.

Public Scholarship

However effective the PPJ is in cultivating habits of excellent public deliberation, if philosophers, activists, policy makers and citizens refuse to write and reflect publicly about the issues with which they are concerned, the wellspring from which PPJ intends to draw its content will run dry. This is no small challenge for many of us in the academy, since it requires and a willingness to subject our ideas the judgment of a much wider public, often at a much earlier stage in their development, than we have historically done.

As Kathleen Fitzpatrick has written: “We too often keep our work as scholars hidden away from the cultural mainstream, pointing toward a pervasive anti-intellectualism that disqualifies the public from engaging with our ideas.” This presumption of anti-intellectualism is a poor rationalization for the continuing insular provincialism of the academy in general and the discipline of Philosophy in particular. Instead, as Fitzpatrick argues, “We must open ourselves up in order to be part of rather than apart from contemporary culture, and in order to do so, we need to expand and rethink the very idea of who our peers are today.”

To expand and rethink who our peers are is in no small part the work of the Public Philosophy Journal. In this regard, the “public” is the first and last concern of the journal. It is the sphere from which its content arises, in which its work is vetted, refined and reviewed, and to which its claims are subject. The PPJ is designed to be public through and through.

Here the difficult work of public scholarship, pursued and developed in public, can be seen to dovetail with and reinforce the importance of practices of public deliberation. The ultimate success of the PPJ will depend upon our ability to create a virtuous circle between public deliberation and public scholarship, between the manner in which the public engages and responds to work by philosophers, activists and policy makers and their willingness to expose their scholarly work to a wide and engaged public. Such a virtuous circle is only possible in a cultivated community of participants willing to risk their ideas and engage one another responsibly. Through this, what Dewey called the “give-and-take of communication,” the Public Philosophy Journal seeks to cultivate a human and humane public scholarly community in a digital age.

References

An overview of developments in the context of public universities can be found in *Peters, Scott J. Engaging Campus and Community: The Practice of Public Scholarship in the State and Land-Grant University System*. Dayton, Ohio: Kettering Foundation Press, 2005. Web. 6 Mar. 2014.

The American Philosophical Association now has a committee devoted to promoting public engagement (publicphilosophy.org/mission.html).

Dewey, John (2008). *The Later Works of John Dewey*, Volume 2, 1925 - 1953: 1925-1927, Essays, Reviews, Miscellany, and the Public and Its Problems. SIU Press, 332.

None of the many open access philosophy journals worldwide (see, www.doaj.org) incorporates all of these dimensions and none leverages open access in a way that so intrinsically links the content and the manner of its production to the mission of the journal. Digital Humanities Now digitalhumanitiesnow.org and its sister publication, The Journal of Digital Humanities journalofdigitalhumanities.org come closest to the spirit of public scholarship the PPJ seeks to embody. Digital Humanities Now uses aggregation, curation, discovery and review to generate articles that appear ultimately in The Journal of Digital Humanities. We intend to integrate some of their practices of curation and aggregation into the workflow of the PPJ.

This report from a 2010 meeting convened at the Pacific Division of the APA can be accessed here (publicphilosophynetwork.ning.com/page/founding-report).

See, for example, the 2011 *Kettering Foundation Report: The Promises and Problems of Public Deliberation* kettering.org/publications.

You can contribute to the PPJ's *Best Practices for Public Scholarly Deliberation* by commenting on the draft we provide here: ppj.matrix.msu.edu/best-practices-for-public-scholarly-deliberation

Fitzpatrick, Kathleen (2011). *Planned Obsolescence*. New York: New York University Press, 17.

Ibid., 17.

Library Science and Textual Transmission in the Online Age: A Fluid Text Model and Proposed Documenting Infrastructure

MacCall, Steven L.

University of Alabama, United States of America

This poster addresses the question: What role can library science and its cooperative organizing methods play within the infrastructure required to facilitate, capture, and preserve the processes involved with the transmission of texts from codex to online? The maturing of online publishing technologies, as evidenced by the present high- rate of innovation in digital scholarly editing, complicates the matter. To deal with this question, library scientists must first make the case to librarians that they are involved in a time of historic change as texts transmit from print books to the online environment and as new types of book and book-like formats emerge online. But more importantly, fundamental research in library science is needed as a basis upon which to develop 21st century professional practice. This poster presents a library science-derived model for the cooperative organizing of the various aspects involved with codex to online textual transmission that is based on the fluid text theory of textual theorist John Bryant. We briefly describe the model and how it provides a basis for an organizing method that relates bibliographical and other evidence centered on a targeted "fluid" text. As a test of the generalizability of the model, we also show how it is extendable to a broader McKenzian conception of "text" in the context of the cooperative organizing of digital special collections materials that themselves are transitioning from analog status to the networked digital environment. We close with speculation on what type of documenting infrastructure might be needed to facilitate such organizing methods including

suggestion for intertextual citation approach suitable for an online publishing environment that is absent the familiar pages and page numbering of the codex.

References

- Bryant, J. L.** (2002). *The Fluid Text: A Theory of Revision and Editing for Book and Screen*. University of Michigan Press.
- McKenzie, D.F.** (1999). *Bibliography and the Sociology of Texts*. New York: Cambridge University Press.

The Inherited Self: Reappraising Literary Cultural Heritage through Digital Methods

Malm, Mats Ulrik

University of Gothenburg, Sweden

Bergenmar, Jenny

University of Gothenburg, Sweden

Kokkinakis, Dimitrios

University of Gothenburg, Sweden

Leonard, Peter

Librarian for Digital Humanities Research, Yale University

This project proposes new approaches to cultural heritage by developing new methods of working with digital texts and by defining appropriate research questions. Our goal is to find ways of turning literature, especially prose fiction, into a site of dynamic research in the humanities and social sciences, rather than merely a passive digital repository.

Our point of departure is the view of cultural heritage as largely intended, or willed, to convey a specific collective memory and identity. This perspective in turn strongly affects the construction of individual identity. From this the project elaborates two main conclusions: 1) In order to fully understand our cultural heritage, it is essential to analyse it *against* the self-understanding of the cultures that produced it — evading or by-passing the structures of literary canon-formation. 2) Focussing on the issue of *identity* is an efficient way of developing methodology and performing analysis.

The project is designed to

- Benefit from a corpus of specially-prepared material where questions of canon formation can be explored through **marginalized and forgotten literary works**
- Develop **new methods of working** with the specific forms of cultural heritage embodied in electronic text databases
- Develop new perspectives and methods through **interdisciplinary exchange and cooperation** on these text databases.

Although the primary material of the pilot project is Swedish, all parts of the project are planned to be generalizable, scalable and relevant to other literary traditions. The main material of the investigations consists of three corpora: The literary works of August Strindberg (based on recently-finalized scholarly editions), the literary works of Selma Lagerlöf (all first editions, established and proofread in collaboration with the scholarly edition), and all original Swedish prose fiction that was first published in the years 1800, 1820, 1840, 1860, 1880 and 1900.

These three corpora offer an apposite opportunity to compare and collate results: Strindberg and Lagerlöf are both canonized, fairly contemporary but entirely different authors: one male and intensely occupied with the societal issues of his day, the other female and developing her own kind of "saga" style, interested in social issues but in a more indirect way. While Strindberg and Lagerlöf belong to Sweden's most renowned and internationally famous authors, the Swedish Prose Fiction database has been constructed in order to evade canonical selection. Comprising all publications that match the criteria, it

offers ways into both mainstream works and those that have been entirely marginalised.

As this project arises from the view of culture as an issue of *identity*, and of cultural heritage as the performative expression of collective memory and identity, the research questions focus on issues of identity: both collective and individual. Since fiction's main means of portraying problems and ideas is the individual character, the studies start out with the individual in order to reach conclusions also about collective identity. The research questions include issues of identity in connection to ethnicity, society, gender and consumption patterns.

The project thus explores and develops different forms of materials, techniques, methods and co-operations, which are to result in new combinations of quantitative and qualitative analysis. In particular, we aim at refining methods of "distant reading", as once proposed by Franco Moretti (Moretti, 2005, 2006), into new approaches that focus on content and context (cf. Jockers, 2013). We use the new tool for sub-corpus topic modeling (STM) designed by Peter Leonard (Leonard and Tangherlini, 2013), which makes it possible to extract topics from a particular work and run against larger materials. We also plan to enhance topic modeling further by adding Named Entity Recognition (NER) and sentiment analysis (cf. Liu, 2010, Maas et al., 2011) to existing systems. NER has been refined, adapted and extended in connection with this project in Kokkinakis and Oelke, 2012, Oelke et al. 2012, Kokkinakis and Malm 2011, 2013; cf. Yang et al., 2011.

At the poster presentation, we will demonstrate materials and techniques on lap-tops.

References

- Jockers, M.** (2013). *Macroanalysis: Digital Methods and Literary History*. Urbana: University of Illinois Press.
- Kokkinakis, D. and Malm, M.** (2011). Character Profiling in 19th Century Fiction. In *Workshop: Language Technologies for Digital Humanities and Cultural Heritage in conjunction with the Recent Advances in Natural Language Processing* (RANLP). Hissar, pp. 70-77.
- Kokkinakis, D. and Oelke, D.** (2012). Men, Women and Gods: Distant Reading in Literary Collections – Combining Visual Analytics with Language Technology. In *Proceedings of the Advances in Visual Methods for Linguistics* (AVML). University of York.
- Kokkinakis D. and Malm M.** (2013). A Macroanalytic View of Swedish Literature using Topic Modeling. In *Proceedings of the Corpus Linguistics*. Andrew Hardie and Robbie Love (eds), Lancaster: UCREL, pp. 144-147.
- Leonard, P. and Tangherlini, T.** (2013). Trawling in the Sea of the Great Unread: Sub-Corpus Topic Modeling and Humanities Research. *Poetics*, 41: 725-749.
- Liu, Bing.** (2010). Sentiment Analysis and Subjectivity. In Indurkhy, N. and Damerau, F. J. (eds), *Handbook of Natural Language Processing*. Boca Raton, Fla: CRC Press, pp. 627-659.
- Maas, A. L., Daly, R. E., Pham P. T., Huang, D., Ng, A. Y. and Potts, C.** (2011). Learning Word Vectors for Sentiment Analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics* (ACL). Portland, pp. 142-150.
- Moretti, F.** (2005). *Graphs, Maps, Trees. Abstract Models for Literary History*. London: Verso.
- Moretti, F. ed.** (2006). *The Novel. History, Geography, and Culture* 1-2. Princeton: Princeton University Press.
- Oelke, D., Kokkinakis, D. and Malm, M.** (2012). Advanced Visual Analytics Methods for Literature Analysis. In *Language Technology for Cultural Heritage, Social Sciences, and Humanities* (LaTeCH), an EACL Workshop. Avignon: Association for Computational Linguistics, pp. 35-44.
- Yang, T., Torget, A. T. and Mihalcea, R.** (2011). Topic Modeling on Historical Newspapers. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*. Portland: Association for Computational Linguistics, pp. 96-104.

Liberate the Text! TypeWright, Cobre, and MapThePage

Mandell, Laura C.

Texas A&M University, United States of America

Heil, Jacob

Texas A&M University, United States of America

Duguid, Timothy

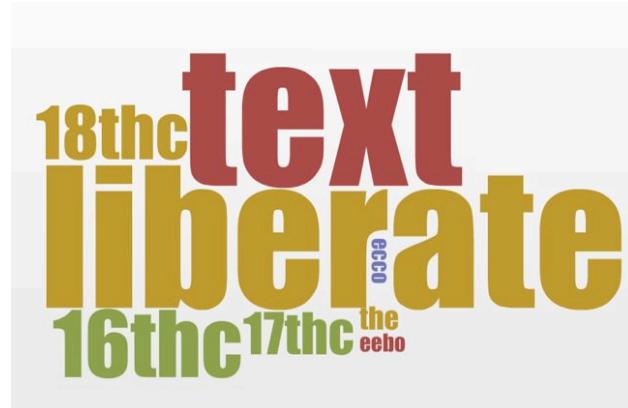
Texas A&M University, United States of America

Grumbach, Elizabeth

Texas A&M University, United States of America

Christy, Matthew

Texas A&M University, United States of America



18thConnect (18thConnect.org) and REKn (the Renaissance English Knowledgebase Project) are two new digital aggregators of early modern digital materials and scholarship, both built on the model of NINES (the Networked Infrastructure for Nineteenth Century Electronic Scholarship, nines.org) (McGann and Nowviskie). Unlike NINES, however, 18thConnect and REKn have two major digital resources for primary materials: Early English Books Online dataset (EEBO) and Eighteenth-Century Collections Online (ECCO), the former owned by ProQuest, the latter by Gale Cengage Learning. Both companies have given us the opportunity to work with the page images in these collections which have been derived, unfortunately, from microfilm. In fact, many of these page images are practically impenetrable to Optical Character Recognition Engines, for reasons having to do both with early print practices during the handpress era and the quality of the images themselves. A consortium of libraries called the Text Creation Partnership (TCP), led by Rebecca Welzenbach at the University of Michigan, has decided to key in, type by hand, one instance of each "title" in the collection. But because metadata for such early texts is notoriously unreliable, the texts not typed may contain some "buried treasures" in the EEBO Collection (Jackson). The TCP has not generated enough money to key ECCO texts: only approximately 1.2% of the 182,000 documents have been typed by the TCP. Texas A&M University has received generous funding from the Andrew W. Mellon Foundation to work on creating Optical Character Recognition training sets for open-source OCR engines that will allow us to mechanically type those page images. However, OCR can only work so well with these images, and so the Early Modern OCR Project (eMOP; emop.tamu.edu) is also building three crowd-source correction tools: TypeWright, Cobre, and MapThePage. This poster will demonstrate these tools, revealing their use in research and teaching. Cobre allows users to consult multiple editions at once in order to corroborate image data. For images that are particularly problematic for the eMOP OCR engines, the AWL editor allows users to edit the bounding boxes in order to produce a more accurate OCR. Finally, within TypeWright users can compare the OCR with the actual images, updating the former as needed. The proposed poster will:

1. Describe the development of each of these three open-source tools.
2. Detail how each tool is needed for creating and verifying OCR data.
3. Demonstrate the online versions of each tool.
4. Show the utility of these tools both inside and outside of the classroom by discussing how they have been used by researchers and students at Texas A&M University.
5. Explain how this improved OCR will then be fed to EEBO and ECCO to enhance scholarship worldwide.
6. Outline the generous contracts we received from ProQuest and Gale: anyone who corrects a text gets to have it and, in effect, liberates it so that it can be full-text searchable, for free, via the 18thConnect and REKn interfaces. Also, scholars are given TEI versions when the text has been freed by their work, enabling them to make digital scholarly editions such as Jess McCarthy's edition of Daniel Defoe's "Hymn to the Pillory" (ahymntothepillory.blogspot.com).

We are about to start a "Liberate the Text!" campaign, and we would like to demonstrate the crowd-sourced correction tools and how to use them as a poster at the DH2014 Conference.

References

- The eMOP Project:* emop.tamu.edu
Jackson, Millie (2008). "Using Metadata to Discover the Buried Treasure in Google Book Search." *Journal of Library Administration* 47.1/2: 165-73.
McGann, Jerome, and Bethany Nowviskie (2005). NINES white paper (PDF, 124kb).

Building the Princeton Prosody Archive

Martin, Meredith

Princeton University, Department of English

Wythoff, Grant

Columbia University, Society of Fellows

Wilson, Meagan

Princeton University, Department of English

Brown, Travis

University of Maryland, Maryland Institute for Technology in the Humanities

1. Introduction

The Princeton Prosody Archive (PPA) is a full-text searchable database of nearly 10,000 digitized texts – comprising 800 million words – on prosody published in English between 1750 and 1923. During the 2012-2013 academic year, a grant from the Mellon Foundation supported the completion of the PPA's first phase. This poster will reflect on the outcomes of the start-up stage, as well as some of the challenges and opportunities the PPA anticipates as the digital collection expands. Conference participants are encouraged to visit prosody.princeton.edu to access the Archive's beta-site.

2. Prosody and Historical Poetics

In the nineteenth century, "prosody" – which refers to both pronunciation and the technicalities of versification – was codified as the fourth section of the grammar book, after "orthography," "etymology," and "syntax." By the early twentieth century, "prosody" referred primarily to versification. In recent years, scholars of English literature have begun questioning the uniformity of poetic terminology, recognizing terms such as "prosody," "meter," "tone," or "rhythm" as culturally determined and fundamentally unstable concepts that have shifted through

the centuries. By turning to historical texts, they are tracing how inherited notions of poetic form developed over time, and in turn, painting a more accurate picture of the evolution of English-language discourse on poetics.

As the field of historical poetics has grown, so too has our access to nineteenth-century materials through online archives. The majority of these digital resources, however, are primarily focused on prose works, and thus both technological and scholarly innovation has been made in the field of prose. The PPA is filling the gap as the only digital archive dedicated to the study of poetics, writ large, and allowing scholars to practice the kind of broad-view historical research the field demands. The PPA aggregates foundational texts in the history of poetics, reviews of these texts, debates about poetics in the public press, and grammar books and poetic handbooks that present contrary definitions and views of poetics so that "big questions" about literary movements and culture can be posed. With this large data-set, we can now ask: How did the changing science of linguistics and increased impulse toward education impact discussions of poetry over time? How often were particular poets used as examples in poetic pedagogy? How, when, and why did certain poetic terms and genres come in and out of use?

3. Methodology

The PPA partnered with Google Books and HathiTrust in 2011, and the collection is currently composed of works digitized by Google and Hathi.[1] In 2013, the PPA began to develop a beta-site, which, though still under development, allows users to browse, search, and correct its content. To best serve its user community, the PPA functions as a freely-available, user-friendly repository, a trusted scholarly reference source, and a creative workspace that enhances traditional scholarly practices and pedagogies while enabling new ones.

3.1 Curation: Google Books and the HathiTrust Digital Library

The PPA's initial corpus was selected from the holdings of the HathiTrust Digital Library. We began by gathering every out-of-copyright text referred to by prosody scholar T.V.F. Brogan in his annotated bibliography English Versification, 1570-1980 that had been digitized.[2] Though the availability of Hathi's digital facsimiles and transcriptions is incredibly valuable, some aspects of their digitization and description present serious technical obstacles to the kinds of analysis the PPA intends to support. The most obvious is that the transcriptions were prepared by a range of Optical Character Recognition (OCR) systems, and few (if any) were hand corrected. Most were digitized as part of the Google Books program, whose OCR tools are not tailored to the vocabulary, orthographic conventions, or typefaces of eighteenth and nineteenth century texts. They were generally unable to capture indentation, italicization, or other formatting, variations in font size, or diacritical marks, not to mention musical notation or non-standard marks.

3.2 OCR and Diacritic Correction: Representing Scansion

When dealing with texts on prosody and versification, accurate representation of diacritics and typographical marks is particularly important. How do you render musical annotation, scansion, line spacing, or iambic markings, for example, into plain text? Because of the focus on notation and the transmission of concepts and terms, particular care must be taken to ensure that these issues do not interfere with (or silently distort) scholarly analysis. To that end, we are developing a model for encoding scansion – the non-textual elements such as musical notation, macron, breve, or other diacritics, including non-standard marks created by the many scholars who attempted to invent prosodic systems in English. Moreover, we will employ the kind of OCR that retains

document coordinates for individual characters whose position on the page often conveys important information.

3.3 Metadata Correction: Scholarly Re-use and Linking Data

The metadata we ingest from HathiTrust also presents challenges. One of the PPA's goals is to allow researchers to trace the development of prosodic discourse across time and place, and the ability to support this functionality depends on consistent and reliable metadata. While the HathiTrust provides the Machine-Readable Cataloging (MARC) records that have been supplied by contributing libraries, the fields indicating the place and date of publication are free text and vary widely in their conventions of encoding. In the PPA's start-up phase, we developed an application that assembles text and metadata from the HathiTrust Digital Library, performs some initial automated correction, and loads the text and metadata into a Drupal 7 installation, where it can be browsed, searched, and corrected by scholars working with the Archive. Corrections to metadata can be credited to registered and authenticated users, and metadata fields can now even be versioned, using Drupal 7's native revision control. In this initial phase, however, these corrections are essentially locked in the Drupal data store; they cannot be returned to the HathiTrust Digital Library or conveniently shared with other scholars working with the same HathiTrust volumes in other contexts. Going forward, the PPA will explore possibilities for enacting a workflow on its own metadata, engaging in the correction of HathiTrust metadata and connecting those corrections to linked data resources by working with the Maryland Institute for Technology in the Humanities and the Foreign Literatures in America project .

3.4 Connecting Prosody Networks: Topic Modeling and Visualization

Topic modeling, and specifically Latent Dirichlet allocation (LDA) has received attention in the digital humanities community over the past several years, in part because it is an unsupervised method – it does not require expensive training material or elaborate encodings – and also because it is relatively robust against textual errors. We have begun experimenting with LDA, not only to return a set of “topics” (which are simply distributions over the vocabulary) that often characterize the semantic and thematic composition of the PPA’s corpus in compelling ways, but also as a means by which we can identify mistranscription, special characters, and even musical notation. We also plan to begin experimenting with visualization tools in the following ways: 1) Plotting temporal and geographical metadata; tools such as Google Earth , MIT’s SIMILE , and Leaflet offer practical and intuitive ways to allow users to navigate temporally and geographically situated data sets interactively – for example, to view a three-dimensional chart on a globe indicating the relative prominence of cities as places of publication while moving a time slider through several centuries; 2) Mapping the documents in the corpus by its topical or lexical spaces; here, each document is represented as a point in a high-dimensional space, where the dimensions of the space are features such as counts or frequencies of individual words or n-grams, or the percentage of words allocated to a particular topic in a topic model; 3) tracking discursive networks by quotation identification and citation extraction; for example, the quotation of exemplars could be represented as a bimodal network, with nodes representing both volumes in the archive and lines of verse, and with edges from the former to the latter indicating instances of quotation.

3.5 Sharing Results

The PPA is committed to providing models so that other digital humanists struggling with the question of how to organize and present their own Hathi collections (in their research or in the classroom). Though these scholars might not be subject area experts in prosody or historical poetics, we would

like to provide enough information that we might navigate unspecialized visitors through the corpus and share ideas about how they might build similar archives themselves.

Notes

[1] We negotiated a Google Distribution Agreement between the Princeton University Library, Princeton Counsel, and HathiTrust that allowed us to access, download, and host all of this data on our own servers. A spreadsheet of all Archive monographs is available online at “Princeton Prosody Archive Database.” The PPA’s four collections can also be accessed through the HathiTrust site. See: 1) “Brogan’s English Versification, 1570-1980” (578 works); 2) “Prosody Archive” (1,308 works); 3) “PPA Subject Search” (6,991 works); and 4) “Graphically/Typographically Unique” (26 works set aside as possessing especially complex page images that would be misread by OCR).

[2] **Brogan, Terry V. F.** *English Versification, 1570-1980: A Reference Guide with a Global Appendix*. Baltimore: Johns Hopkins University Press, 1981.

TEI Customization for encoding paratexts in spanish printed books (XV-XVIII)

Martos Pérez, María Dolores

mdmartos@flog.uned.es
UNED

Baranda Leturio, Nieves

nbaranda@flog.uned.es
UNED

Marín Pina, M^a Carmen

mmarin@unizar.es
Universidad de Zaragoza

1. Introduction

1.1. Overview

El propósito de este póster es proporcionar una demostración del proyecto de etiquetado y edición digital con TEI de los paratextos en obras impresas de escritoras españolas desde el siglo XV al XVIII de BIESES (www.bieses.net). La información contextual que aportan los paratextos sobre sus autoras, el contexto de producción, las motivaciones que impulsaron su escritura, su recepción o su proceso de edición exigen una visión dinámica de la obra literaria mediante su representación digital, no como un producto hecho (*work*) sino como un acto de producción de sentido (*act*) que se reactualiza en cada cultura o época y en cada lector (“text in performance” [Bryant, 2002]¹).

1.2. Methodology

Partiendo de la experiencia de toda una serie de proyectos de investigación dedicados a la escritura femenina (Brown University Project, African American Women Writers of the 19th Century, British Women Romantic Poet, 1789-1832, Women writer Resource Pages, UNiversidad de Chicago, The Orlando Project, Victorian Women Writers Project, Indiana UNiversity, Emory Women Writer Project)², estamos fijando esquemas de etiquetado, a partir de un thesaurus de “valores” aplicados a los atributos que definan la posición de las escritoras en el campo literario, sus actitudes ante la escritura y las valoraciones que

sus contemporáneos vierten en estos textos preliminares o epilogales.

El codificado de este tipo de textos invierte el esquema tradicional de TEI, que privilegia el desarrollo del "body" frente al "front" y "back", dado que los paratextos ocupan un espacio "marginal" en la obra literaria: portada, preliminares (antes del cuerpo de la obra) y epílogo (después del cuerpo de la obra); además de otro tipo de informaciones paratextuales que se pueden intercalar en el cuerpo del texto y otro tipo de glosas o anotaciones marginales, que también hay que codificar.

2. Getting Started

En nuestro proyecto prescindimos del "body", que queda reducido a un enlace a un texto digitalizado o a una breve mención del inicio o final de la obra literaria en cuestión, y codificamos únicamente los textos reservados a los módulos del "front" y "back" [Burnard-Rahtz 2004: 5-8]³.

Por analogía de temas e intereses seguimos muy de cerca el modelo de customización [Bauman y Flander 2004]⁴ de *Women Writer Project* (Brown University), que usa el "ODD language" (véase: www.wwp.brown.edu)⁵.

En esta primera fase la customización se está centrándolo en la eliminación de elementos innecesarios y en la definición y establecimiento de valores para atributos, centrados en la caracterización de los enunciadores de estos textos (escritoras y destinatarios) y en las valoraciones sobre el proceso de escritura. A medida que el proceso de customización avance puede resultar necesario cambiar la estructura interna de algunos elementos para representar la realidad textual de los paratextos.

Por todo ello creamos que, siguiendo los modelos citados de proyectos que ya tiene una amplia andadura, la customización a través de TE IODD que proponemos para los paratextos de obras literarias españolas escritas entre el siglo XV-XVIII, centrada en los elementos de *front* y *back* puede resultar interesante para otros investigadores y proyectos que atiendan a la literatura femenina de este período.

References

1. **Bryant, Jonh** (2002), *The fluid Text: A Theory of Revision and Editing for Book and Screen*, Ann Arbor, University of Michigan Press
2. digital.lib.ucdavis.edu/projects/bwrep/ ; meet.tge-adonis.fr/projet/roles-et-pouvoirs-des-femmes-au-xvie-siecle-dans-la-france-de-louest ; webapp1.dlib.indiana.edu/vwwp/welcome.do ;jsessionid=9135BDD19390863752F291EAEA63E22E y <https://wiki.dlib.indiana.edu/display/vwwp/Home>; www.womenwriters.nl/index.php/Prefaces ; womenwriters.library.emory.edu/ ; www.artscrn.ualberta.ca/orlando/ ; www.wwp.brown.edu/http://digital.nypl.org/schomburg/writers_aa19/
3. **Lou Burnard y Sebastian Rahtz**, *Relax NG with Son of ODD*, Extreme Markup Language, 2004: citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.124.7139&rep=rep1&type=pdf
4. **Bauman, Syd y Flander, Julia**, *Odd customizations*, Extreme Markup Language, 2004: conferences.idealliance.org/extreme/html/2004/Bauman01/EML2004Bauman01.html
5. www.wwp.brown.edu/research/publications/guide/html/customization.html

Kenny, Julia

julia.kenny90@gmail.com

Università di Pisa

Di Pietro, Chiara

dipi.chiara@gmail.com

Università di Pisa

The Digital Vercelli Book project¹ aims at creating a digital edition of the Vercelli Book manuscript (MS CXVII, Biblioteca Capitolare di Vercelli, Italy), a fundamental document for the Old English language and literature area of studies. After the transcription of all the texts was completed using the TEI schemas², the researchers lacked a tool that would allow to publish a diplomatic edition of the former together with the digitized manuscript images.

Many tools have been created in order to publish digital editions based on XML-encoded texts and meet the various needs of users. Some software tools (such as TEI Boilerplate) allow to publish TEI XML files directly in modern browsers. Even though this method is very light and user friendly, it is both unsuitable for publishing very large files and limited to XSLT 1.0; the typical output is a single web page for each XML document, which is not appropriate for the publication of image-based editions.

This led us to the design and implementation of EVT (Edition Visualization Technology), a lightweight software for that specific purpose. What started as a project, whose goal was the digital edition of a specific manuscript, has grown well beyond that, to the point of being available and usable as a general purpose publishing tool.

EVT can be used to create image-based web editions with different edition levels starting from a multi level encoded text. It is built on open and standard web technologies, such as HTML, CSS and Javascript, to ensure that it will be working on all the most recent web browsers. During the development of the software we have constantly tested it on different cases of study (for instance the NZTEC corpus), in order to verify that the code was actually compatible with the different kinds of TEI P5 encoding.

The general architecture of the software is modular, so that any component can be easily upgraded or replaced. To overcome the limitations hinted above the data processing is carried out by means of a static transformations with XSLT 2.0. EVT's processing system uses a collection of stylesheets to divide the XML file into smaller portions (each corresponding to individual pages of the manuscript), and for each of these parts it creates as many output files as requested by the file settings (for instance, diplomatic and diplomatic-interpretative). On the image side, the system offers a ready-to-use set of instruments (such as a magnifying lens, a general zoom, hot spots) to fully take advantage of manuscripts' scans. One of the most important tools available in EVT is the image-text link: if the XML files make use of the <facsimile> element and provide the coordinates that identify the lines in manuscript scans, the EVT building system will automatically create clickable areas on the image, which will be handled thanks to CSS and Javascript interaction.

Edition Visualization Technology: a simple tool to publish digital editions and digital facsimiles

Masotti, Raffaele

raffaele.masotti@gmail.com

Università di Pisa

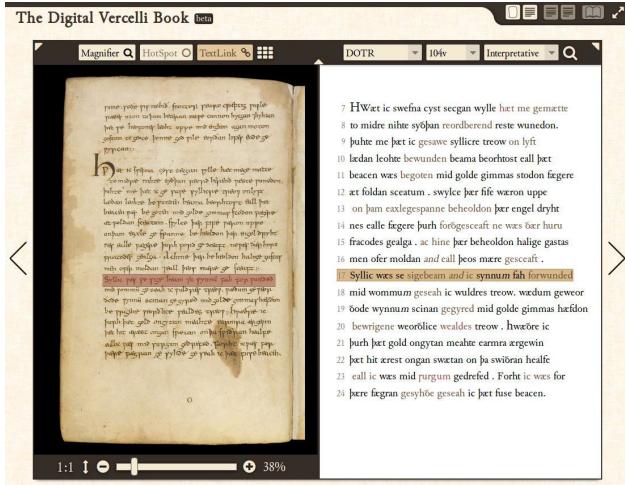


Fig. 1: Main view

The most important feature is that the code is working as a client-only infrastructure, so that the user may immediately make use of it, without requiring any server software installation. This is apparently a limitation, since a client-server architecture would offer many more features, but thanks to the processing done with the XSLT 2.0 processor in the initial phase of the building process the system is also ready to perform complex operations (such as textual searches) which generally rely on interpreted languages and a server back-end.

What has been described up to this point constitutes the core of our project, that is a complex and modular collection of XSLT scripts coupled with text and image browsing tools. The modules are designed in such a way that the scholar will be able to add his own stylesheets to manage the different levels of the edition, and this will not influence the other parts of the system, f.i. HTML generation. Applying the XSLT scripts usually requires the use of specific editors (such as Oxygen), which may be not the user's favorite tool and add one more step to the process. This is why we decided to create EVT Interface: a Java user interface that allows to avoid loading the XML files in an editor and makes directly use of the open source versions of the Saxon processor to apply XSLT modules to the TEI XML files.

Our main objective is to create a flexible, modular, stand-alone package which is also freely redistributable.

We are now facing several research questions (UI evolution, search engine, embedded transcription, critical edition) that we hope to address during the next few months.

References

- Vercelli Book Digitale.** vbd.humnet.unipi.it/ (accessed on March 2014).
- Burnard, L., and S. Bauman** (eds.) *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. 2.6.0. www.tei-c.org/P5/.

Our Marathon: The Boston Bombing Digital Archive

McGrath, Jim

mcgrath.ja@husky.neu.edu
Northeastern University

Peaker, Alicia

peaker.a@husky.neu.edu
Northeastern University

In May 2013, graduate students and faculty members at Northeastern University's NULab for Texts, Maps, and Networks began work on **Our Marathon: The Boston**

Bombing Digital Archive (www.northeastern.edu/marathon). Motivated by the interest Northeastern's students and faculty members displayed in sharing their stories about the 2013 Boston Marathon bombings with one another, Our Marathon is an ambitious endeavor to create a central repository of stories and content related to the event and its aftermath. In the same ways that the September 11 Digital Archive and The Hurricane Digital Memory Bank utilized crowdsourcing to gather material, Our Marathon has reached out to a wide range of individuals (within and beyond the Boston community) to collect stories, photos, social media, and oral histories.

The project is invested in the role stories and community-building play in responding to traumatic events: it encourages members of the public to share the stories they may have already told about the Marathon on sites like Facebook, Twitter, and Instagram, and it provides a site for individuals to process and to reflect on the event in a variety of genres. Like its archival predecessors, Our Marathon is committed to creating a digital forum that has value to individuals in the immediate present and to researchers and other interested parties in the decades to come. Using the Omeka web-publishing platform, the archive has been attentive to the long-term preservation standards favored by archivists and university librarians, gathering and updating metadata on its items with DublinCore standards in mind. That being said, the project also seeks to convince an audience beyond these academic contexts that the work of digital humanists is also valuable to them.

The ongoing work of building the Our Marathon digital archive has raised several questions that may be of interest to other digital humanists. How do investments in academic and non-academic audiences inform a digital archive's content, interface, and accessibility? How can digital archivists productively collaborate across disciplines at home institutions, with students and scholars at other academic institutions, with community organizations, and with media partners? What are some of the challenges of creating a digital archive about a traumatic event shortly after its occurrence and within close proximity to many of the communities and areas directly affected by that event? What are the advantages and disadvantages of crowdsourced initiatives? How can digital archives compellingly represent and catalogue material like Facebook statuses and Tweets for both researchers and the public? What steps can digital humanists take to ensure long-term preservation of their digital projects? We also encourage attendees to navigate the archive and to share their own stories with Our Marathon.

Speaking in Code

Nowviskie, Bethany

University of Virginia Library, United States of America

Rochester, Eric

University of Virginia Library, United States of America

Graham, Wayne

Boggs, Jeremy

University of Virginia Library, United States of America

McClure, David

University of Virginia Library, United States of America

Bailey, Scott

University of Virginia Library, United States of America

Many digital humanities projects are collaborations involving not only traditionally-employed scholars, but also technologists and administrative staff. These groups all have their own traditions of discourse, evolving methodologies, metrics of success, sources of recognition, and avenues for promotion. Any point of difference can become a source of useful and creative energy, or a nexus of misunderstanding and conflict.

The gaps in communication that these differences open can be deep and broad, even among humanities-trained software developers and the scholars with whom they collaborate. Much

(not all) knowledge advances in software development through hands-on, journeyman learning experiences and the iterative, often-collaborative development of built objects and systems. Much (not all) knowledge advances in humanities scholarship through fixed and fluid kinds of academic discourse: referential, prosy, often agonistic. Divisions can exist in style and practice, even when the subjects and objects of humanities inquiry are the same. What approaches might bridge the gaps between tacit knowledge exchange and the writing of humanities theory and interpretation? Can we move past an historical moment in the academy, in which the onus seems to be almost entirely placed on archivally and theoretically trained humanities scholars to become tech-savvy *digital humanist*, in order to build a concomitant sense of momentum, responsibility, and opportunity in our community of DH software engineers? Can we build greater community itself, just by making a space in which such problems are addressed?

In early November 2013, the Scholars' Lab at the University of Virginia Library hosted an event called "Speaking in Code"—a two-day, high-level summit for approximately 30 advanced humanities software developers. Participants were selected on the basis of their demonstrated experience in digital humanities software development, their interest in advancing solutions to the problems raised by the summit, and the disciplinary and cultural diversity they promised bring to the conversation. The Scholars' Lab team also made a clear and explicit call for participation by developers who are women, people of color, queer/LGBT, or otherwise under-represented among the ranks of digital humanities programmers, and we were well pleased at the response. (For instance, 12 of our 30 participants identified as women and 4 volunteered that they are LGBT.) Besides the unusual level of gender diversity at the event, we believe it was unique in other ways. Our summit—supported by generous grants from the National Endowment for the Humanities and the UVa Library—was the first focused meeting to address scholarly and social implications of tacit knowledge exchange in the digital humanities.

First-day discussions at "Speaking in Code" (led by Bethany Nowviskie, William J. Turkel, Stéfan Sinclair, Mia Ridge, and Hugh Cayless) addressed core problems and activities in humanities computing. These included: physical and digital embodiment, and our unspoken understandings of them, as made evident in code; how best to take advantage of moments of fruitful rupture between design and development phases in DH work; challenges in crafting humanistic models for the representation and procedural analysis of human language; and methods by which developers and metadata specialists grapple with other kinds of ambiguity, or "messy understandings," in cultural heritage information. The second day started with concrete project pitches, responding to day-one conversation and offered by participants in lightning rounds. This was followed by work on some of the projects that were pitched, conducted in small groups, with an eye both toward making immediate interventions in the field and seeding longer-term, collaborative undertakings. Our DH 2014 poster will present outcomes from "Speaking in Code."

The underlying question of our summit was this: how might we—at a moment when scholarly interest in humanities computing is growing by leaps and bounds—bring longstanding technical conversations into more open, inclusive humanities discourse? "Speaking in Code" foregrounded the intellectual dimensions of DH craftsmanship but—importantly, unusually, and we think as a necessary first step to fostering discussion in venues legible and friendly to scholars and developers alike—we started with a meeting *organized and conducted on software developers' own terms*.

References

- Collins, Harry M.** (2010). *Tacit and Explicit Knowledge*. Chicago: The University of Chicago Press.
- Liu, Alan** (2012). "The state of the digital humanities: A report and a critique." *Arts and Humanities in Higher Education*. Feb/Apr 2012 11: 8-41.
- Mattern, Shannon** (2011). "Revisiting Craft I: Teaching the Connections Between Thinking and Making." 18 August 2011. <<http://www.wordsinspace.net/wordpress/2011/08/18/revisiting-craft-i-teaching-the-connections-between-thinking-and-making/>>
- Mattern, Shannon** (2011). "Revisiting Craft II: Tools of Craftsmanship." 18 August 2011. <<http://www.wordsinspace.net/wordpress/2011/08/18/revisiting-craft-2-tools-and-methods-of-craftsmanship/>>
- Oram, Andy, and Greg Wilson** (2007). *Beautiful Code: Leading Programmers Explain How They Think* (Theory in Practice). O'Reilly Media, Inc..
- Polanyi, Michael** (1958). *Personal Knowledge: Towards a Post-Critical Philosophy*. Chicago: University of Chicago Press.
- Polanyi, Michael** (1966). *The Tacit Dimension*. Chicago: University of Chicago Press, 1966. Scholars' Lab. Speaking in Code. Scholars' Lab. 2013. Web. 1 Nov 2013. <<http://codespeak.scholarslab.org/>>
- Wardrip-Fruin** (2009), Noah. *Expressive processing: digital fictions, computer games, and software studies*. Cambridge, Mass.: MIT Press
- Wood, Nicola, et al** (2009). "A Tacit Understanding: The Designer's Role in Capturing and Passing on the Skilled Knowledge of Craftsmen." Working Paper, Art and Design Research Centre, Sheffield Hallam University. <<http://www.archive.org/details/ATacitUnderstanding>>

La Hiperedição Dos Panfletos De Eulálio Motta: Edición Filológica Y Cultura Digital, Retos De Un Nuevo Tiempo

Nunes Barreiros, Patrício

particiobarreiros@hotmail.com
Universidade Estadual De Feira De Santana, Brazil

Introducción: situando el debate

El advenimiento de la informática trajo consigo cambios expresivos en el modo como la sociedad se relaciona con la escrita, estableciendo una nueva organización de los modos de producción, transmisión, preservación y usos de los textos. Del mismo modo, las disciplinas que utilizan el texto como objeto de estudio también fueron obligadas a repensaren sus prácticas investigativas. En el ámbito de la Crítica Textual, por ejemplo, los filólogos están utilizando la tecnología digital para diseñar nuevos modelos de edición y para poner en práctica un nuevo concepto de filología. Pero, trasladar textos de la cultura manuscrita y/o impresa al medio digital nos es una tarea simple y requiere pericia filológica (Lucía Megías, 2010).

Los textos no son hechos solamente de una secuencia alfanumérica. En su naturaleza están implicados también los códigos bibliográficos y contextuales que contribuyen para su entendimiento (Bornstein, 2001; McGann, 1992, 1991). Los diferentes modos como los textos ganaron forma, fueron distribuidos, leídos, comercializados, almacenados, utilizados, etc. hacen parte de su historia primordial como lo ha señalado tantas veces Roger Chartier (2001; 2002). Por eso, a la hora de editar textos que tuvieron su existencia en la cultura material se necesita conservar, además de los códigos lingüísticos, los contextuales y bibliográficos.

Una secuencia de letras y palabras (entendido como texto) puede ser reproducida en diferentes soportes y tipografías. Un texto manuscrito en cursiva humanística escrito en un pergamo, con letras capitulares ornadas, con filigranas y dibujos, puede ser trasladado al impreso tipográfico y después al digital. Supongamos que el único testimonio manuscrito de ese texto tenga huellas de quemaduras en sus bordas, contenga correcciones y anotaciones marginales. Si al editar el texto, el filólogo da atención solamente al establecimiento del código lingüístico, todos los significados que se pueden extraer del soporte, de la técnica de escritura, del *design caligráfico* y tipográfico se pierden o son señalados en largas notas,

aparatos críticos o introducciones exhaustivas que los lectores no suelen leer.

Al ingresar en el mundo digital, un texto oriundo de la cultura material pasa a compartir de una lógica, bastante diferente de la cual él fue creado y construyó su historia. De ese modo, el filólogo que asume la responsabilidad de preparar una edición digital de textos impresos o manuscritos, tiene la responsabilidad de preservar, lo más posible, los códigos inherentes al texto para no alejarlo de su socio-historia. Pero esta no es una tarea sencilla y exige diálogo con diferentes disciplinas.

La Hiperedición de los panfletos de Eulálio Motta

Eulálio Motta (Mundo Novo-Ba, Brasil, 1907-1988) organizó un archivo personal que contiene miles de documentos diversos (libros, manuscritos de sus obras, cuadernos con anotación diaria, cartas, fotografías, colecciones de periódicos, etc.). Entre 1930 y 1988 él escribió panfletos, hojas sueltas impresas en tipografías manuales que eran distribuidas en la ciudad de Mundo Novo, estado de Bahia, Brasil. Por ser una literatura efímera solamente el autor conservó los panfletos en su acervo personal.

Al identificar los panfletos en el archivo del escritor, se observó que los impresos estaban atados a otros documentos tales como cartas, fotografías, postales, manuscritos, libros, etc. Editar los panfletos sin relacionarlos a tales documentos sería alejarlos de su socio-historia. Además de esto, las técnicas utilizadas para la impresión de los panfletos corresponden a la historia cultural de las prácticas de escrita del Sertão de Bahia, siendo imprescindible preservar los códigos bibliográficos, el *diesign* y el *layout* de los impresos.

Por lo tanto, se diseñó un modelo de edición digital que valorase la documentación paratextual (organizado en el *dossiê arquivístico*) y conjugase en un solo sitio transcripción semidiplomática, facsímile, fotografías y hasta video. Según Jerome McGinn, este tipo de edición digital que reúne hipertexto, video, iconografía y textos dinámicos, se configura como una hipermedia que es en verdad una hiperedición.

Para la hiperedición se necesitó:

1. organizar el acervo del escritor: describir, catalogar, elaborar un banco de datos digital con el catálogo del acervo y hacer el inventario;
2. organizar los dossiers arquivísticos de todos los panfletos;
3. digitalizar los panfletos y los documentos de los dossiers arquivísticos;
4. editor el video;
5. transcribir todos los documentos textuales utilizados en la edición;
6. preparar las ediciones críticas;
7. elaborar las imágenes de la transcripción sobrepuerta al facsímile;
8. preparar una versión para imprimir de todos los textos de la edición;
9. estructurar el banco de datos del Google Drive para la descripción linear de los textos; 1
10. hacer una descripción paleográfica de los panfletos;
11. elaborar el layout y el design gráfico de las páginas de la edición digital;
12. proceder a la etiquetaje del texto y a la edición de la página (utilizando Personal Home Page - PHP y HTML5, a través del editor Adobe Dreamweaver 5);
13. desarrollar funciones de script a partir de JavaScript.



Fig. 1: Página Inicial de la edición

www.eulaliomotta.com.br

A partir de la página inicial el usuario puede acceder a las ediciones de los panfletos pulsando en las imágenes de los impresos, organizados cronológicamente y con movimiento. Además de esto, hay muchas informaciones contextuales sobre la vida del escritor (álbum, entrevista, etc.). El usuario puede dejar su recado y puede observar los avisos dejados por el editor.

Al pie de la página hay las informaciones sobre los créditos y el aviso sobre los derechos autorales.



Fig. 2: Perfil del Escritor

Hay varios perfiles del escritor: poeta, periodista, político, pasquineiro, álbum, bibliografía.



Fig. 3: El álbum de Fotografías

El usuario puede ver a un conjunto de fotografías organizadas por el editor. Son documentos que pertenecen al archivo del escritor y trae y trae informaciones en el verso.

This screenshot displays a detailed view of a pamphlet page. The page contains dense text in Portuguese. On the left side, there is a sidebar with various menu options and links. Numbered callouts point to specific features: 1) Facsímile; 2) Transcripción del texto sobre el facsímile; 3) Transcripción linear del texto; 4) Transcripción con link; 5) Crea una versión para imprimir; 6) Zoom hasta 100%; 7) Información paleográfica del texto (textura del papel, gramatura, color, lisura, dimensiones, etc.); 8) Dossiê arquivístico con los documentos del acervo del escritor que se relacionan con el panfleto editado, todos los ítems del dossiê son editados del mismo modo del panfleto en el centro, con las mismas opciones de edición; 9) Información histórica sobre el documento; 10) Índice temático y por título de los panfletos.

Fig. 4: La edición de un Panfleto

1. Facsímile;
2. Transcripción del texto sobre el facsímile;
3. Transcripción linear del texto;
4. Transcripción con link;
5. Crea una versión para imprimir;
6. Zoom hasta 100%;
7. Información paleográfica del texto (textura del papel, gramatura, color, lisura, dimensiones, etc.);
8. Dossiê arquivístico con los documentos del acervo del escritor que se relacionan con el panfleto editado, todos los ítems del dossiê son editados del mismo modo del panfleto en el centro, con las mismas opciones de edición;
9. Información histórica sobre el documento;
10. Índice temático y por título de los panfletos.

This screenshot shows a zoomed-in view of a historical document. The text is partially visible, showing words like 'que impo'. The interface includes a vertical toolbar on the right with options for transcription, printing, and other functions. A copyright notice at the bottom states: 'Copyright © 2013 Pátria Nova Sistemas, todos os direitos reservados. O conteúdo dessa site está protegido pela lei de direitos autorais, não é permitido a reprodução ou reutilização de nenhum dos documentos em questão. Acesso a partir da página oficial 'O Pasquineiro da Roça' do programa de Pós-Graduação em Letras e Linguística da Universidade Federal de Santa Catarina, Florianópolis, 2013, sob a supervisão do Prof. Dr. César Augusto Tadeu de Oliveira.'

Fig. 5: Visualización en Zoom

Todos los textos o imágenes visualizadas pueden ser aumentadas hasta 100%.

This screenshot displays a detailed view of a handwritten letter. The text is in cursive script. On the left side, there is a sidebar with various menu options and links. Numbered callouts point to specific features: 1) Facsímile; 2) Transcripción del texto sobre el facsímile; 3) Transcripción linear del texto; 4) Transcripción con link; 5) Crea una versión para imprimir; 6) Zoom hasta 100%; 7) Información paleográfica del texto (textura del papel, gramatura, color, lisura, dimensiones, etc.); 8) Dossiê arquivístico con los documentos del acervo del escritor que se relacionan con el panfleto editado, todos los ítems del dossiê son editados del mismo modo del panfleto en el centro, con las mismas opciones de edición; 9) Información histórica sobre el documento; 10) Índice temático y por título de los panfletos.

Fig. 6: Edición de una carta > dossiê arquivístico de un pafleto

This screenshot displays a transcription of a letter. The text is in cursive script and is being transcribed into a linear format. On the left side, there is a sidebar with various menu options and links. Numbered callouts point to specific features: 1) Facsímile; 2) Transcripción del texto sobre el facsímile; 3) Transcripción linear del texto; 4) Transcripción con link; 5) Crea una versión para imprimir; 6) Zoom hasta 100%; 7) Información paleográfica del texto (textura del papel, gramatura, color, lisura, dimensiones, etc.); 8) Dossiê arquivístico con los documentos del acervo del escritor que se relacionan con el panfleto editado, todos los ítems del dossiê son editados del mismo modo del panfleto en el centro, con las mismas opciones de edición; 9) Información histórica sobre el documento; 10) Índice temático y por título de los panfletos.

Fig. 7: Ejemplo de transcripción con link

Conclusión

La informática y su utilización por las humanidades exigen una revisión metodológica de las ciencias tradicionales, pero

señala el valor de estas en el entorno digital. En el campo de la filología, por ejemplo, se nota la necesidad de utilización de los conocimientos especializado de la crítica textual, de la paleografía, de la historia cultural de escrita, de la diplomática, etc. asociados al diseño gráfico y a la informática. Por todos ganan con estos cambios y diálogos.

Para la edición de los panfletos de Eulálio Motta, la hiperedición cumplió satisfactoriamente el intento de expresar los códigos lingüísticos, bibliográficos y contextuales, como se pretendía desde el inicio del proyecto.

References

- Barreiros, Patrício.** www.eulaliomotta.com.br .
- Barreiros, Patrício Nunes.** (2013). *O Pasquineiro Da Roça: Edição Dos Panfletos De Eulálio Motta.* 386F. Tese (Doutorado Em Letras) - Instituto De Letras - Universidade Federal Da Bahia, Salvador
- Bornstein, George.** (2001). *Material Modernism, The Politics Of The Page.* New York: Cambridge University Press.
- Chartier, Roger.** (2001). *Cultura Escrita, Literatura E História: Conversas De Roger Chartier Com Carlos Aguirre Anaya, Jesús Anaya Rosique, Daniel Goldin E Antônio Saborit.* Porto Alegre: Artmed
- Chartier, Roger.** (2002). *Os Desafios Da Escrita.* Tradução De Fulvia M. L. Moretto. São Paulo: Unesp
- Lucía Megías, José Manuel.** (2006). *De Las Bibliotecas A Las Plataformas De Conocimiento (Notas Sobre El Futuro Del Texto En La Era Digital).* In: Santiago, Ramón; Valenciano, Ana; Iglesias, Silvia. *Tradiciones Discursivas, Edición De Textos Orales Y Escritos.* Madrid: Editorial Complutense.
- Lucía Megías, José Manuel.** (2007). *Hacia Nuevos Paradigmas Textuales* (Edición Y Difusión De Los Textos Literarios En El Siglo XXI). Madrid: Universidad Complutense De Madrid.
- Lucía Megías, José Manuel.** (2008). *La Informática Humanística: Una Puerta Abierta Para Los Estudios Medievales En El Siglo XXI.* Revista De Poética Medieval, N. 20, Madrid, P. 163-185
- Lucía Megías, José Manuel.** (2010). *Reflexiones En Torno A Las Plataformas De Edición Digital: El Ejemplo De La Celestina.* In: Poalini, Devid. (Coord.). *De Ninguna Cosa Es Alegre Posesión Sin Compañía, Estudios Celestinos Y Medievales En Honor Del Profesor Joseph Thomas Snow.* Tomo I. New York: Seminário Hispánico De Estudios Medievales, P. 226-251
- Lucía Megías, José Manuel.** (2012). *Elogio Del Texto Digital, Claves Para Interpretar El Nuevo Paradigma.* Madrid: Fórcola
- Marigno, Emmanuel.** (2012). *Edición Digital Y Edición Impresa: La Obra Satírica De Quevedo.* In: Bravo, Federico. *Desafíos Y Perspectivas De La Edición Digital.* Villa María: Euvim, P. 23-34
- Mcgann, Jerome.** (1991). *The Textual Condition.* Princeton: Princeton University Press
- Mckenzie, Donald Francis.** (2005). *Bibliografía Y Sociología De Los Textos.* Madrid: Akal.

The Digitization of Hmong Sacred Texts

Ogden, Mitchell Paul

University of Wisconsin-Stout, US

A remote religious community, called Ee Nbee Mee Noo (Ib Npis Mis Nus), in northern Thailand has recently shared their extensive collection of hand-written religious manuscripts with researchers following extensive collaboration and community work. Interested in making these foundational sacred texts more accessible across the global Hmong diaspora, the community leaders sought the cooperation of scholars to digitize the texts in preparation for eventual publication. Rich in historical and

religious content, these nine 200-page volumes are written in the community's own sacred and obscure Puaj Txwm alphabet of the Hmong language.

The scope of our project includes the development of a set of digital tools that will perform OCR (built on the Tesseract engine), error-check based on phonological rules of the Hmong language, and encode the text (according to TEI guidelines) to make the text searchable and indexable for future distribution as electronic texts. In addition, the project will create a transliteration tool that will enable Hmong texts to be transliterated across the dozen or more orthographies used throughout the diaspora, thus facilitating the exchange and investigation of Hmong language texts across diasporic histories and geographies.

This poster will feature the project's tools and process, emphasizing the solutions to obstacles including a hand-written manuscript and the complicated politics of ethnic minority religious movements in Thailand. The poster will provide context for these sacred texts and their obscure alphabet, overview the digitization process, and provide examples of the work and the innovative solutions developed by our research team.

Digital Humanities as Vocation: Possibilities for Undergraduate Education

Ogden, Mitchell Paul

University of Wisconsin-Stout, United States of America

The concept of *vocation* has continually evolved—from its Catholic origins through the long history of Western education and through the cultural transformation of the Industrial Revolution. In the midst of the Information Revolution, vocation continues to evolve, accommodating emerging trades and the day-to-day operations of participants in a digital information economy. We've seen, for example, that the pace of innovation and technological change has destabilized the model of fixed-skills training, demanding instead a rapid and recurrent retraining and retooling of the workforce—and not just for production- or entry-level employees, but for employees across the labor spectrum.

In our contemporary economic moment, there is an increasing demand for information workers who rely upon applied conceptual knowledge over their ability to simply use existing tools to accomplish existing tasks. Flexible and versatile thinkers—ready to advance with the vanguard—are the order of the day. It is not difficult to make the argument that the skill sets of digital humanists are well matched for this information economy: familiarity with software development and tool-building, intellectually-intensive cross-disciplinary collaboration, appreciation of interface design and visualization, among many others. As the field of digital humanities continues to expand, there is a steady increase of undergraduate programs that provide opportunities for students to learn and apply DH methodologies, build DH tools, and cultivate DH skill sets. Examining the rise of undergraduate digital humanities work on American liberal arts campuses, Bryan Alexander and Rebecca Frost Davis predict that those undergraduate programs will be sending their students into digital humanities graduate programs (Kindle Locations 9938-9939). But we know that not every student will have interest or opportunity to pursue graduate studies or to enter the academic DH industry. So what sort of professional opportunities will await graduates of these proliferating undergraduate programs? Can we think about digital humanities as a vocation—that rich professional calling that is both spiritually and economically sustaining and sustainable? Put another way, is there digital humanities work outside the academy?

Scholarship of teaching and learning in the digital humanities remains scant. In one of the few treatments of the subject,

Stephen Brier raises the big question "Where's the pedagogy?" His project traces the legacy of "innovative pedagogy" at the City University of New York. That pedagogical legacy includes providing access to higher education to a large population of working-class students in New York and the surrounding region—students entrenched in the vocational identities of their families and communities. Brier's narrative of innovation is one that sees the continual infusion of technology and digital tools to aid student learning that eventually culminates in the pathbreaking digital humanities initiatives housed at CUNY. But while undergraduate students across the curriculum use and engage digital tools, serious DH work is reserved for graduate students and faculty who address teaching and learning as a meta-discursive component of their research. There is a strong and viable connection between DH research and teaching and learning, but digital humanities remains aloof from undergraduate experience and distant from the grounding in vocation.

Blackwell and Martin explore the practice—increasingly common among DH researchers—of involving undergraduate students as research assistants in DH projects. As Blackwell and Martin suggest, these experiences facilitate a mastery of humanities inquiry—classics in their case—and point to a fertile area to see enhancement of teaching and learning. As important as such undergraduate research opportunities are, however, they remain—in most cases—extracurricular activities meant to enhance a student's education experience rather than define it.

This short paper is interested in exploring the possibility of developing digital humanities as an undergraduate curriculum—especially in a context that privileges a legacy of manual and vocational training. It considers the case of the University of Wisconsin–Stout—a polytechnic university with a historical identity of manual and vocational preparation and an institutional focus on job placement. In 2010, the Professional Communication and Emerging Media program at UW–Stout revised its curriculum to include a new concentration in digital humanities. Different from both research universities and small liberal arts colleges, this polytechnic campus has a teaching-centered mission that is rich with programs that complement digital humanities, including entertainment design, game design & development, media production, computer science, graphic arts, and applied social sciences. As a new program in this novel environment, we are just beginning to see how undergraduate students respond to the digital humanities curriculum, how the first cohorts of students prepare to enter the labor market, and how they shape their own vocational identities.

Such a program raises a number of serious questions. As humanists and digital humanists, how un/comfortable are we with framing any part of our field as vocational training? As the field expands, will we be prepared to think more strategically—more vocationally—about job markets beyond academia? On one hand, vocational thinking goes against the grain of the liberal education sensibilities carried by many members of the DH community. On the other, the rapid expansion of interest and support for our field raises the question: Why? What are the forces—larger than academia's craving for intellectually fashionable movements—that drive this boom? Among many possible answers to this question, could it be that our methodologies and modes of inquiry are recognized to be well-aligned with the Zeitgeist of our information economy? Might digital humanities be a "new" field that has emerged in response to broader labor and economic forces and not just—as some continue to believe—a technological makeover of a has-been discipline?

Put another way, might digital humanities be viable as a progressive training model for the labor pool of the surging information economy? Can the digital humanities be a twenty-first century vocation outside of academia?

References

- 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. Minneapolis: U of Minnesota P.

Blackwell, Christopher and Thomas R. Martin (2009). "Technology, Collaboration, and Undergraduate Research." *Digital Humanities Quarterly*. 3.1: n. pag. Web. 30 Oct. 2013.

Brier, Stephen (2012). "Where's the Pedagogy? The Role of Teaching and Learning in the Digital Humanities." Debates in the Digital Humanities. Ed. Matthew Gold. Minneapolis: U of Minnesota P.

Modeling Melville's Reading: Editing Marginalia in TEI, Topic Modeling Reading and Influence

Ohge, Christopher M.

christopherohge@gmail.com

University of California at Berkeley, US

I propose to contribute a poster at the Digital Humanities 2014 Convention on my recent initiative with Melville's Marginalia Online (www.melvillesmarginalia.org). In addition to creating an interactive, reader-friendly digital bibliography of the books Herman Melville was known to have owned, borrowed, and consulted, the editors of the project have been editing digital editions of the surviving books that contain Melville's marginalia. The next phase for the project is to mark up the marginalia files in TEI P5, as well as thinking of new ways to represent the individuality as well as the totality of Melville's reading practices. I have taken responsibility for this phase, first by collecting the OCR'd files of Melville's editions, then by finding a way to mark up the marginalia. One challenge of the project is that it uses the "coordinate capture" tool (created by Matt Cohen at the University of Texas at Austin and the Walt Whitman Archive), which creates XML coordinates corresponding to the image file of each and every aspect of the book (from spine to covers to individual pages). However, these coordinate-captured XML files cannot be manipulated, which complicates the task of marking up the marginalia up in TEI. Furthermore, currently there exists no standard in the TEI Guidelines for marking up marginalia (especially as those kinds of "notes" correspond to specific places in the text while not being written in a linear fashion in most cases). Yet another question remains about Melville's reading: how can we better understand not only his reading practices, but also quantitatively understand how his reading affected his published work? Countless studies have elucidated Melville's sources (both with solid research methods and conjecture), but I propose to include a topic model (using Mallet) on Melville's Marginalia Online that will "read" Melville's reading in ways that will change the way we think about how the works he read influenced him. No longer must we guess how his reading influenced him; topic modeling will let us read the library of his entire life, and apply that information to his published writings. Gathering data through topic modeling allows Melville scholars a new way of studying literary influence. In the poster session, I look forward to reporting on the TEI markup of the marginalia files in order to show other attendees working on authors' libraries how best to accomplish this task, as well as to demonstrate a custom algorithm for topic modeling large literary corpora relating to authorial influence.

Large-scale text analysis through the HathiTrust Research Center

Organisciak, Peter

organis2@illinois.edu

Graduate School of Library and Information Science, University of Illinois

Bhattacharyya, Sayan

sayan@illinois.edu

Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign

Auvil, Loretta

lauvil@illinois.edu

University of Illinois at Urbana-Champaign

Plale, Beth

plale@cs.indiana.edu

Data To Insight Center, Indiana University Bloomington

Downie, J. Stephen

jdownie@illinois.edu

Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign

Introduction

Digitization of text and tools for making sense of it are enabling digital humanists to perform ever-larger exploration and inference, from early work analyzing the style of a relatively small set of works or a single writer^{1 2}, to the modern-day practice of “distant reading” entire eras.^{3 4} However, “research, even in the digital age, is... limited by the materials that scholars can readily and reliably access”.⁵ Such factors include copyright, availability of materials, cost of infrastructure, and/or the technical capabilities demanded of the researcher tend to limit the access that digital humanists have, in practice, to the texts with which they would like to work.

Overview

The HathiTrust Research Center is a collaborative effort formed in 2011 through a partnership between Indiana University, the University of Illinois at Urbana-Champaign, and the HathiTrust, to meet the challenge of dealing with massive amounts of digital text that digital humanists confront when they perform “distant reading.” We will present recent progress by the HTRC in addressing this ongoing challenge.

HTRC aims to support the natural investigative process of researchers who want to perform text analytics on the HathiTrust corpus by running the analytics algorithms “close to the data,” even when content restrictions do not allow actual human-reading level (“consumptive”) access to the text. A popular parallel to this higher-level exploration of digital corpora is the Google Ngram Viewer⁶. HTRC’s tools maintain a conceptually similar non-consumptive separation from the content, while allowing more control over the content being looked at.

The HathiTrust public domain corpus consists of an online repository comprising a comprehensive body of published works drawn from the collections of over sixty participating major research institutions and libraries. Digital humanists can access digitized works through the HTRC via two different levels: the production-system and the sandbox. The production level of the HTRC provides access to the public domain HathiTrust corpus (a mix of works digitized by Google and other digitization projects), secured to comply with the restrictions on the content use. In contrast, the sandbox level of the HTRC provides a more open level of access, to a smaller corpus consisting of 250,000 volumes which do not have any known copyright restrictions. Building on top of the SEASR tools⁷, both systems support analytics such as topic modeling, tag clouds, entity extraction, spellcheck reports, and the Dunning’s log-likelihood statistic on the distribution of text. They also include functionalities such as a Marc downloader and a word frequencies. The intention is for researchers to design algorithms on the more open sandbox system and then submit the algorithms on the production system. A metadata and data API exist on the sandbox system as well as on the production system for accessing metadata and token counts.

Methodology

In this poster, we focus on the questions that humanities scholars can address using the HTRC’s front-end tools. Specifically, HTRC offers a workset builder for searching the HathiTrust collection and creating collections of texts (“worksets”), and a portal for analyzing such worksets through a simple web interface.

The screenshot shows the HTRC Workset Builder interface. At the top, there are links for 'Log Out [Peter Organisciak]', 'Selected Items (100)', 'Manage Worksets', and a 'Portal' link. Below this is a 'RESEARCH CENTER' logo. The main area is titled 'Selected Items' and shows a list of 100 items. The first item is '11. Noon.' with details: Title: Noon., Language: English, Published: 1900. The second item is '12. Poems / by Elizabeth Watts.' with details: Title: Poems / by Elizabeth Watts., Author: Watts, Elizabeth, Language: English, Published: 1900. There are buttons for 'Create/Update Workset' and 'Clear all'. A checkbox labeled 'Selected' is checked next to the first item. Navigation buttons 'Previous' and 'Next' are at the bottom.

Fig. 1: Results are collected for a workset of Late 19c Poetry.

The screenshot shows the HTRC Portal interface. The top navigation bar includes 'HTRC Portal', 'Home', 'About', 'Workers', 'Algorithms', 'Results', 'Help', and a 'Signed-in as: organisciak' link. Below this is a 'Job Details' section for a job titled '11. Noon.'. It lists various parameters: 'Name: Meandre_TagCloud_with_Cleaning', 'Description: This analysis performs token counts with some additional text cleaning and displays the most frequent tokens as a tag cloud. Loads each page of each volume from HTRC. Removes the first and last line of each page. Joins hyphenated words that occur at the end of the line. Performs lowercase transformation of text. Removes all tokens that don't consist of alphabetic characters. Uses the replacement rules (learned from our usage of Google Ngrams data) to clean OCR errors, normalize to British spelling and normalize for period spelling. Filters stop words. Counts the tokens remaining for all volumes and displays the top 200 tokens in a tag cloud. NOTE: The volume limit is 1000.', 'Version: 1.1', 'Author: Loretta Auvil', and 'Please input Job Name: (required) Late19cPoetry'. Below this is a 'Please select a collection for analysis' dropdown set to 'Late19cPoetry @organisciak'. Further down are sections for 'replacement_rules_url' (with options for 'TextCorrections' and 'TextCommon') and 'stopwords_list_english_url' (with options for 'TextCommon_stopwords' and 'TextCommon_stopwords').

Fig. 2: Setting up an algorithm to run on the workset.

The screenshot shows the HTRC Portal interface with the 'View Results' tab selected. The results for the job 'Late19cPoetry' are displayed as a tag cloud. The most prominent words include 'long', 'american', 'poem', 'edition', 'great', 'life', 'poems', 'work', 'york', 'thou', 'london', 'year', 'full', 'the', 'poet', 'love', 'printed', 'man', 'god', 'series', 'faith', 'form', 'true', 'edges', 'death', 'blank', 'ire', 'quot', 'de', 'work', 'calf', 'vol', 'heart', 'york', '8th', 'great', 'calf', 'street', 'red', 'age', 'notes', 'war', 'vo', 'rare', 'thou', 'li', 'mol', 'london', 'day', 'year', 'full', 'ppl', 'life', 'poems', 'work', 'the', 'poet', 'love', 'printed', 'man', 'god', 'series', 'faith', 'form', 'true', 'edges', 'death', 'blank', 'ire', 'quot', 'de', 'work', 'calf', 'vol', 'heart', 'york', '8th', 'great', 'calf', 'street', 'red', 'age', 'notes', 'war', 'vo', 'rare', 'thou', 'li', 'mol', 'london', 'day', 'year', 'full', 'ppl', 'life', 'poems', 'work', 'the', 'poet', 'love', 'printed', 'man', 'god', 'series', 'faith', 'form', 'true', 'edges', 'death', 'blank', 'ire', 'quot', 'de', 'work', 'calf', 'vol', 'heart', 'york', '8th', 'great', 'calf', 'street', 'red', 'age', 'notes', 'war', 'vo', 'rare', 'thou', 'li', 'mol', 'london', 'day', 'year', 'full', 'ppl', 'life', 'poems', 'work', 'the', 'poet', 'love', 'printed', 'man', 'god', 'series', 'faith', 'form', 'true', 'edges', 'death', 'blank', 'ire', 'quot', 'de', 'work', 'calf', 'vol', 'heart', 'york', '8th', 'great', 'calf', 'street', 'red', 'age', 'notes', 'war', 'vo', 'rare', 'thou', 'li', 'mol', 'london', 'day', 'year', 'full', 'ppl', 'life', 'poems', 'work', 'the', 'poet', 'love', 'printed', 'man', 'god', 'series', 'faith', 'form', 'true', 'edges', 'death', 'blank', 'ire', 'quot', 'de', 'work', 'calf', 'vol', 'heart', 'york', '8th', 'great', 'calf', 'street', 'red', 'age', 'notes', 'war', 'vo', 'rare', 'thou', 'li', 'mol', 'london', 'day', 'year', 'full', 'ppl', 'life', 'poems', 'work', 'the', 'poet', 'love', 'printed', 'man', 'god', 'series', 'faith', 'form', 'true', 'edges', 'death', 'blank', 'ire', 'quot', 'de', 'work', 'calf', 'vol', 'heart', 'york', '8th', 'great', 'calf', 'street', 'red', 'age', 'notes', 'war', 'vo', 'rare', 'thou', 'li', 'mol', 'london', 'day', 'year', 'full', 'ppl', 'life', 'poems', 'work', 'the', 'poet', 'love', 'printed', 'man', 'god', 'series', 'faith', 'form', 'true', 'edges', 'death', 'blank', 'ire', 'quot', 'de', 'work', 'calf', 'vol', 'heart', 'york', '8th', 'great', 'calf', 'street', 'red', 'age', 'notes', 'war', 'vo', 'rare', 'thou', 'li', 'mol', 'london', 'day', 'year', 'full', 'ppl', 'life', 'poems', 'work', 'the', 'poet', 'love', 'printed', 'man', 'god', 'series', 'faith', 'form', 'true', 'edges', 'death', 'blank', 'ire', 'quot', 'de', 'work', 'calf', 'vol', 'heart', 'york', '8th', 'great', 'calf', 'street', 'red', 'age', 'notes', 'war', 'vo', 'rare', 'thou', 'li', 'mol', 'london', 'day', 'year', 'full', 'ppl', 'life', 'poems', 'work', 'the', 'poet', 'love', 'printed', 'man', 'god', 'series', 'faith', 'form', 'true', 'edges', 'death', 'blank', 'ire', 'quot', 'de', 'work', 'calf', 'vol', 'heart', 'york', '8th', 'great', 'calf', 'street', 'red', 'age', 'notes', 'war', 'vo', 'rare', 'thou', 'li', 'mol', 'london', 'day', 'year', 'full', 'ppl', 'life', 'poems', 'work', 'the', 'poet', 'love', 'printed', 'man', 'god', 'series', 'faith', 'form', 'true', 'edges', 'death', 'blank', 'ire', 'quot', 'de', 'work', 'calf', 'vol', 'heart', 'york', '8th', 'great', 'calf', 'street', 'red', 'age', 'notes', 'war', 'vo', 'rare', 'thou', 'li', 'mol', 'london', 'day', 'year', 'full', 'ppl', 'life', 'poems', 'work', 'the', 'poet', 'love', 'printed', 'man', 'god', 'series', 'faith', 'form', 'true', 'edges', 'death', 'blank', 'ire', 'quot', 'de', 'work', 'calf', 'vol', 'heart', 'york', '8th', 'great', 'calf', 'street', 'red', 'age', 'notes', 'war', 'vo', 'rare', 'thou', 'li', 'mol', 'london', 'day', 'year', 'full', 'ppl', 'life', 'poems', 'work', 'the', 'poet', 'love', 'printed', 'man', 'god', 'series', 'faith', 'form', 'true', 'edges', 'death', 'blank', 'ire', 'quot', 'de', 'work', 'calf', 'vol', 'heart', 'york', '8th', 'great', 'calf', 'street', 'red', 'age', 'notes', 'war', 'vo', 'rare', 'thou', 'li', 'mol', 'london', 'day', 'year', 'full', 'ppl', 'life', 'poems', 'work', 'the', 'poet', 'love', 'printed', 'man', 'god', 'series', 'faith', 'form', 'true', 'edges', 'death', 'blank', 'ire', 'quot', 'de', 'work', 'calf', 'vol', 'heart', 'york', '8th', 'great', 'calf', 'street', 'red', 'age', 'notes', 'war', 'vo', 'rare', 'thou', 'li', 'mol', 'london', 'day', 'year', 'full', 'ppl', 'life', 'poems', 'work', 'the', 'poet', 'love', 'printed', 'man', 'god', 'series', 'faith', 'form', 'true', 'edges', 'death', 'blank', 'ire', 'quot', 'de', 'work', 'calf', 'vol', 'heart', 'york', '8th', 'great', 'calf', 'street', 'red', 'age', 'notes', 'war', 'vo', 'rare', 'thou', 'li', 'mol', 'london', 'day', 'year', 'full', 'ppl', 'life', 'poems', 'work', 'the', 'poet', 'love', 'printed', 'man', 'god', 'series', 'faith', 'form', 'true', 'edges', 'death', 'blank', 'ire', 'quot', 'de', 'work', 'calf', 'vol', 'heart', 'york', '8th', 'great', 'calf', 'street', 'red', 'age', 'notes', 'war', 'vo', 'rare', 'thou', 'li', 'mol', 'london', 'day', 'year', 'full', 'ppl', 'life', 'poems', 'work', 'the', 'poet', 'love', 'printed', 'man', 'god', 'series', 'faith', 'form', 'true', 'edges', 'death', 'blank', 'ire', 'quot', 'de', 'work', 'calf', 'vol', 'heart', 'york', '8th', 'great', 'calf', 'street', 'red', 'age', 'notes', 'war', 'vo', 'rare', 'thou', 'li', 'mol', 'london', 'day', 'year', 'full', 'ppl', 'life', 'poems', 'work', 'the', 'poet', 'love', 'printed', 'man', 'god', 'series', 'faith', 'form', 'true', 'edges', 'death', 'blank', 'ire', 'quot', 'de', 'work', 'calf', 'vol', 'heart', 'york', '8th', 'great', 'calf', 'street', 'red', 'age', 'notes', 'war', 'vo', 'rare', 'thou', 'li', 'mol', 'london', 'day', 'year', 'full', 'ppl', 'life', 'poems', 'work', 'the', 'poet', 'love', 'printed', 'man', 'god', 'series', 'faith', 'form', 'true', 'edges', 'death', 'blank', 'ire', 'quot', 'de', 'work', 'calf', 'vol', 'heart', 'york', '8th', 'great', 'calf', 'street', 'red', 'age', 'notes', 'war', 'vo', 'rare', 'thou', 'li', 'mol', 'london', 'day', 'year', 'full', 'ppl', 'life', 'poems', 'work', 'the', 'poet', 'love', 'printed', 'man', 'god', 'series', 'faith', 'form', 'true', 'edges', 'death', 'blank', 'ire', 'quot', 'de', 'work', 'calf', 'vol', 'heart', 'york', '8th', 'great', 'calf', 'street', 'red', 'age', 'notes', 'war', 'vo', 'rare', 'thou', 'li', 'mol', 'london', 'day', 'year', 'full', 'ppl', 'life', 'poems', 'work', 'the', 'poet', 'love', 'printed', 'man', 'god', 'series', 'faith', 'form', 'true', 'edges', 'death', 'blank', 'ire', 'quot', 'de', 'work', 'calf', 'vol', 'heart', 'york', '8th', 'great', 'calf', 'street', 'red', 'age', 'notes', 'war', 'vo', 'rare', 'thou', 'li', 'mol', 'london', 'day', 'year', 'full', 'ppl', 'life', 'poems', 'work', 'the', 'poet', 'love', 'printed', 'man', 'god', 'series', 'faith', 'form', 'true', 'edges', 'death', 'blank', 'ire', 'quot', 'de', 'work', 'calf', 'vol', 'heart', 'york', '8th', 'great', 'calf', 'street', 'red', 'age', 'notes', 'war', 'vo', 'rare', 'thou', 'li', 'mol', 'london', 'day', 'year', 'full', 'ppl', 'life', 'poems', 'work', 'the', 'poet', 'love', 'printed', 'man', 'god', 'series', 'faith', 'form', 'true', 'edges', 'death', 'blank', 'ire', 'quot', 'de', 'work', 'calf', 'vol', 'heart', 'york', '8th', 'great', 'calf', 'street', 'red', 'age', 'notes', 'war', 'vo', 'rare', 'thou', 'li', 'mol', 'london', 'day', 'year', 'full', 'ppl', 'life', 'poems', 'work', 'the', 'poet', 'love', 'printed', 'man', 'god', 'series', 'faith', 'form', 'true', 'edges', 'death', 'blank', 'ire', 'quot', 'de', 'work', 'calf', 'vol', 'heart', 'york', '8th', 'great', 'calf', 'street', 'red', 'age', 'notes', 'war', 'vo', 'rare', 'thou', 'li', 'mol', 'london', 'day', 'year', 'full', 'ppl', 'life', 'poems', 'work', 'the', 'poet', 'love', 'printed', 'man', 'god', 'series', 'faith', 'form', 'true', 'edges', 'death', 'blank', 'ire', 'quot', 'de', 'work', 'calf', 'vol', 'heart', 'york', '8th', 'great', 'calf', 'street', 'red', 'age', 'notes', 'war', 'vo', 'rare', 'thou', 'li', 'mol', 'london', 'day', 'year', 'full', 'ppl', 'life', 'poems', 'work', 'the', 'poet', 'love', 'printed', 'man', 'god', 'series', 'faith', 'form', 'true', 'edges', 'death', 'blank', 'ire', 'quot', 'de', 'work', 'calf', 'vol', 'heart', 'york', '8th', 'great', 'calf', 'street', 'red', 'age', 'notes', 'war', 'vo', 'rare', 'thou', 'li', 'mol', 'london', 'day', 'year', 'full', 'ppl', 'life', 'poems', 'work', 'the', 'poet', 'love', 'printed', 'man', 'god', 'series', 'faith', 'form', 'true', 'edges', 'death', 'blank', 'ire', 'quot', 'de', 'work', 'calf', 'vol', 'heart', 'york', '8th', 'great', 'calf', 'street', 'red', 'age', 'notes', 'war', 'vo', 'rare', 'thou', 'li', 'mol', 'london', 'day', 'year', 'full', 'ppl', 'life', 'poems', 'work', 'the', 'poet', 'love', 'printed', 'man', 'god', 'series', 'faith', 'form', 'true', 'edges', 'death', 'blank', 'ire', 'quot', 'de', 'work', 'calf', 'vol', 'heart', 'york', '8th', 'great', 'calf', 'street', 'red', 'age', 'notes', 'war', 'vo', 'rare', 'thou', 'li', 'mol', 'london', 'day', 'year', 'full', 'ppl', 'life', 'poems', 'work', 'the', 'poet', 'love', 'printed', 'man', 'god', 'series', 'faith', 'form', 'true', 'edges', 'death', 'blank', 'ire', 'quot', 'de', 'work', 'calf', 'vol', 'heart', 'york', '8th', 'great', 'calf', 'street', 'red', 'age', 'notes', 'war', 'vo', 'rare', 'thou', 'li', 'mol', 'london', 'day', 'year', 'full', 'ppl', 'life', 'poems', 'work', 'the', 'poet', 'love', 'printed', 'man', 'god', 'series', 'faith', 'form', 'true', 'edges', 'death', 'blank', 'ire', 'quot', 'de', 'work', 'calf', 'vol', 'heart', 'york', '8th', 'great', 'calf', 'street', 'red', 'age', 'notes', 'war', 'vo', 'rare', 'thou', 'li', 'mol', 'london', 'day', 'year', 'full', 'ppl', 'life', 'poems', 'work', 'the', 'poet', 'love', 'printed', 'man', 'god', 'series', 'faith', 'form', 'true', 'edges', 'death', 'blank', 'ire', 'quot', 'de', 'work', 'calf', 'vol', 'heart', 'york', '8th', 'great', 'calf', 'street', 'red', 'age', 'notes', 'war', 'vo', 'rare', 'thou', 'li', 'mol', 'london', 'day', 'year', 'full', 'ppl', 'life', 'poems', 'work', 'the', 'poet', 'love', 'printed', 'man', 'god', 'series', 'faith', 'form', 'true', 'edges', 'death', 'blank', 'ire', 'quot', 'de', 'work', 'calf', 'vol', 'heart', 'york', '8th', 'great', 'calf', 'street', 'red', 'age', 'notes', 'war', 'vo', 'rare', 'thou', 'li', 'mol', 'london', 'day', 'year', 'full', 'ppl', 'life', 'poems', 'work', 'the', 'poet', 'love', 'printed', 'man', 'god', 'series', 'faith', 'form', 'true', 'edges', 'death', 'blank', 'ire', 'quot', 'de', 'work', 'calf', 'vol', 'heart', 'york', '8th', 'great', 'calf', 'street', 'red', 'age', 'notes', 'war', 'vo', 'rare', 'thou', 'li', 'mol', 'london', 'day', 'year', 'full', 'ppl', 'life', 'poems', 'work', 'the', 'poet', 'love', 'printed', 'man', 'god', 'series', 'faith', 'form', 'true', 'edges', 'death', 'blank', 'ire', 'quot', 'de', 'work', 'calf', 'vol', 'heart', 'york', '8th', 'great', 'calf', 'street', 'red', 'age', 'notes', 'war', 'vo', 'rare', 'thou', 'li', 'mol', 'london', 'day', 'year', 'full', 'ppl', 'life', 'poems', 'work', 'the', 'poet', 'love', 'printed', 'man', 'god', 'series', 'faith', 'form', 'true', 'edges', 'death', 'blank', 'ire', 'quot', 'de', 'work', 'calf', 'vol', 'heart', 'york', '8th', 'great', 'calf', 'street', 'red', 'age', 'notes', 'war', 'vo', 'rare', 'thou', 'li', 'mol', 'london', 'day', 'year', 'full', 'ppl', 'life', 'poems', 'work', 'the', 'poet', 'love', 'printed', 'man', 'god', 'series', 'faith', 'form', 'true', 'edges', 'death', 'blank', 'ire', 'quot', 'de', 'work', 'calf', 'vol', 'heart', 'york', '8th', 'great', 'calf', 'street', 'red', 'age', 'notes', 'war', 'vo', 'rare', 'thou', 'li', 'mol', 'london', 'day', 'year', 'full', 'ppl', 'life', 'poems', 'work', 'the', 'poet', 'love', 'printed', 'man', 'god', 'series', 'faith', 'form', 'true', 'edges', 'death', 'blank', 'ire', 'quot', 'de', 'work', 'calf', 'vol', 'heart', 'york', '8th', 'great', 'calf', 'street', 'red', 'age', 'notes', 'war', 'vo', 'rare', 'thou', 'li', 'mol', 'london', 'day', 'year', 'full', 'ppl', 'life', 'poems', 'work', 'the', 'poet', 'love', 'printed', 'man', 'god', 'series', 'faith', 'form', 'true', 'edges', 'death', 'blank', 'ire', 'quot', 'de', 'work', 'calf', 'vol', 'heart', 'york', '8th', 'great', 'calf', 'street', 'red', 'age', 'notes', 'war', 'vo', 'rare', 'thou', 'li', 'mol', 'london', 'day', 'year', 'full', 'ppl', 'life', 'poems', 'work', 'the', 'poet', 'love', 'printed', 'man', 'god', 'series', 'faith', 'form', 'true', 'edges', 'death', 'blank', 'ire', 'quot', 'de', 'work', 'calf', 'vol', 'heart', 'york', '8th', 'great', 'calf', 'street', 'red', 'age', 'notes', 'war', 'vo', 'rare', 'thou', 'li', 'mol', 'london', 'day', 'year', 'full', 'ppl', 'life', 'poems', 'work', 'the', 'poet', 'love', 'printed', 'man', 'god', 'series', 'faith', 'form', 'true', 'edges', 'death', 'blank', 'ire', 'quot', 'de', 'work', 'calf', 'vol', 'heart', 'york', '8th', 'great', 'calf', 'street', 'red', 'age', 'notes', 'war', 'vo', 'rare', 'thou', 'li', 'mol', 'london', 'day', 'year', 'full', 'ppl', 'life', 'poems', 'work', 'the', 'poet', 'love', 'printed', 'man', 'god', 'series', 'faith', 'form', 'true', 'edges', 'death', 'blank', 'ire', 'quot', 'de', 'work', 'calf', 'vol', 'heart', 'york', '8th', 'great', 'calf', 'street', 'red', 'age', 'notes', 'war', 'vo', 'rare', 'thou', 'li', 'mol', 'london', 'day', 'year', 'full', 'ppl', 'life', 'poems', 'work', 'the', 'poet', 'love', 'printed', 'man', 'god', 'series', 'faith', 'form', 'true', 'edges', 'death', 'blank', 'ire', 'quot', 'de', 'work', 'calf', 'vol', 'heart', 'york', '8th', 'great', 'calf', 'street', 'red', 'age', 'notes', 'war', 'vo', 'rare', 'thou', 'li', 'mol', 'london', 'day', 'year', 'full', 'ppl', 'life', 'poems', 'work', 'the', 'poet', 'love', 'printed', 'man', 'god', 'series', 'faith', 'form', 'true', 'edges', 'death', 'blank', 'ire', 'quot', 'de', 'work', 'calf', 'vol', 'heart', 'york', '8th', 'great', 'calf', 'street', 'red', 'age', 'notes', 'war', 'vo', 'rare', 'thou', 'li', 'mol', 'london', 'day', 'year', 'full', 'ppl', 'life', 'poems', 'work', 'the', 'poet', 'love', 'printed', 'man', 'god', 'series', 'faith', 'form', 'true', 'edges', 'death', 'blank', 'ire', 'quot', 'de', 'work', 'calf', 'vol', 'heart', 'york', '8th', 'great', 'calf', 'street', 'red', 'age', 'notes', 'war', 'vo', 'rare', 'thou', 'li', 'mol', 'london', 'day', 'year', 'full', 'ppl', 'life', 'poems', 'work', 'the', 'poet', 'love', 'printed', 'man', 'god', 'series', 'faith', 'form', 'true', 'edges', 'death', 'blank', 'ire', 'quot', 'de', 'work', 'calf', 'vol', 'heart', 'york', '8th', 'great', 'calf', 'street', 'red', 'age', 'notes', 'war', 'vo', 'rare', 'thou', 'li', 'mol', 'london', 'day', 'year', 'full', 'ppl', 'life', 'poems', 'work', 'the', 'poet', 'love', 'printed', 'man', 'god', 'series', 'faith', 'form', 'true', 'edges', 'death', 'blank', 'ire', 'quot', 'de', 'work', 'calf', 'vol', 'heart', 'york', '8th', 'great', 'calf', 'street', 'red', 'age', 'notes', 'war', 'vo', 'rare', 'thou', 'li', 'mol', 'london', 'day', 'year', 'full', 'ppl', 'life', 'poems', 'work', 'the', 'poet', 'love', 'printed', 'man', 'god', 'series', 'faith', 'form', 'true', 'edges', 'death', 'blank', 'ire', 'quot', 'de', 'work', 'calf', 'vol', 'heart', 'york', '8th', 'great', 'calf', 'street', 'red', 'age', 'notes', 'war', 'vo', 'rare', 'thou', 'li', 'mol', 'london', 'day', 'year', 'full', 'ppl', 'life', 'poems', 'work', 'the', 'poet', 'love', 'printed', 'man', 'god', 'series', 'faith', 'form', 'true', 'edges', 'death', 'blank', 'ire', 'quot', 'de', 'work', 'calf', 'vol', 'heart', 'york', '8th', 'great', 'calf', 'street', 'red', 'age', 'notes', 'war', 'vo', 'rare', 'thou', 'li', 'mol', 'london', 'day', 'year', 'full', 'ppl', 'life', 'poems', 'work', 'the', 'poet', 'love', 'printed', 'man', 'god', 'series', 'faith', 'form', 'true', 'edges', 'death', 'blank', 'ire', 'quot', 'de', 'work', 'calf', 'vol', 'heart', 'york', '8th', 'great', 'calf', 'street', 'red', 'age', 'notes', 'war', 'vo', 'rare', 'thou', 'li', 'mol', 'london', 'day', 'year', 'full', 'ppl', 'life', 'poems', 'work', 'the', 'poet', 'love', 'printed', 'man', 'god', 'series', 'faith', 'form', 'true', 'edges', 'death', 'blank', 'ire', 'quot', 'de', 'work', 'calf', 'vol', 'heart', 'york', '8th', 'great', 'calf', 'street', 'red', 'age', 'notes', 'war', 'vo', 'rare', 'thou', 'li', 'mol', 'london', 'day', 'year', 'full', 'ppl', 'life', 'poems', 'work', 'the', 'poet', 'love', 'printed', 'man', 'god', 'series', 'faith', 'form', 'true', 'edges', 'death', 'blank', 'ire', 'quot', 'de', 'work', 'calf', 'vol', 'heart', 'york', '8th', 'great', 'calf', 'street', 'red', 'age', 'notes', 'war', 'vo', 'rare', 'thou', 'li', 'mol', 'london', 'day', 'year', 'full', 'ppl', 'life', 'poems', 'work', 'the', 'poet', 'love', 'printed', 'man', 'god', 'series', 'faith', 'form', 'true', 'edges', 'death', 'blank', 'ire', 'quot', 'de', 'work', 'calf', 'vol', 'heart', 'york', '8th', 'great', 'calf', 'street', 'red', 'age', 'notes', 'war', 'vo', 'rare', 'thou', 'li', 'mol', 'london', 'day', 'year', 'full', 'ppl', 'life', 'poems', 'work', 'the', 'poet', 'love', 'printed', 'man', 'god', 'series', 'faith', 'form', 'true', 'edges', 'death', 'blank', 'ire', 'quot', 'de', 'work', 'calf', 'vol', 'heart', 'york', '8th', 'great', 'calf', 'street', 'red', 'age', 'notes', 'war', 'vo', 'rare', 'thou', 'li', 'mol', 'london', 'day', 'year', 'full', 'ppl', 'life', 'poems', 'work', 'the', 'poet', 'love', 'printed', 'man', 'god', 'series', 'faith', 'form', 'true', 'edges', 'death', 'blank', 'ire', 'quot', 'de', 'work', 'calf', 'vol', 'heart', 'york', '8th', 'great', 'calf', 'street', 'red', 'age', 'notes', 'war', 'vo', 'rare', 'thou', 'li', 'mol', 'london', 'day', 'year', 'full', 'ppl', 'life', 'poems', 'work', 'the', 'poet', 'love', 'printed', 'man', 'god', 'series', 'faith', 'form', 'true', 'edges', 'death', 'blank', 'ire', 'quot', 'de', 'work', 'calf', 'vol', 'heart', 'york', '8th', 'great', 'calf', 'street', 'red', 'age', 'notes', 'war', 'vo', 'rare', 'thou', 'li', 'mol', 'london', 'day', 'year', 'full', 'ppl', 'life', 'poems', 'work', 'the', 'poet', 'love', 'printed', 'man', 'god', 'series', 'faith', 'form', 'true', 'edges', 'death', 'blank', 'ire', 'quot', 'de', 'work', 'calf', 'vol', 'heart', 'york', '8th', 'great', 'calf', 'street', 'red', 'age', 'notes', 'war', 'vo', 'rare', 'thou', 'li', 'mol', 'london', 'day', 'year', 'full', 'ppl', 'life', 'poems', 'work', 'the', 'poet', 'love', 'printed', 'man', 'god', 'series', 'faith', 'form', 'true', 'edges', 'death', 'blank', 'ire', 'quot', 'de', 'work', 'calf', 'vol', 'heart', 'york', '8th', 'great', 'calf', 'street', 'red', 'age', 'notes', 'war', 'vo', 'rare', 'thou', 'li', 'mol', 'london', 'day', 'year', 'full', 'ppl', 'life', 'poems', 'work', 'the', 'poet', 'love', 'printed', 'man', 'god', 'series', 'faith', 'form', 'true', 'edges', 'death', 'blank', 'ire', 'quot', 'de', 'work', 'calf', 'vol', 'heart', 'york', '8th', 'great', 'calf', 'street', 'red', 'age', 'notes', 'war', 'vo', 'rare', 'thou', 'li', 'mol', 'london', 'day', 'year', 'full', 'ppl', 'life', 'poems', 'work', 'the', 'poet', 'love', 'printed', 'man', 'god', 'series', 'faith', 'form', 'true', 'edges', 'death', 'blank', 'ire', 'quot', 'de', 'work', 'calf', 'vol', 'heart', 'york', '8th', 'great', 'calf', 'street', 'red', 'age', 'notes', 'war', 'vo', 'rare', 'thou', 'li', 'mol', 'london', 'day', 'year', 'full', 'ppl', 'life', 'poems', 'work', 'the', 'poet', 'love', 'printed', 'man', 'god', 'series', 'faith', 'form', 'true', 'edges', 'death', 'blank', 'ire', 'quot', 'de', 'work', 'calf', 'vol', 'heart', 'york', '8th', 'great', 'calf', 'street', 'red', 'age', 'notes', 'war', 'vo', 'rare', 'thou', 'li', 'mol', 'london', 'day', 'year', 'full', 'ppl', 'life', 'poems', 'work', 'the', 'poet', 'love', 'printed', 'man', 'god', 'series', 'faith', 'form', 'true', 'edges', 'death', 'blank', 'ire', 'quot', 'de', 'work', 'calf', 'vol', 'heart', 'york', '8th', 'great', 'calf', 'street', 'red', 'age', 'notes', 'war', 'vo', 'rare', 'thou', 'li', 'mol', 'london', 'day', 'year', 'full', 'ppl', 'life', 'poems', 'work', 'the', 'poet', 'love', 'printed', 'man', 'god', 'series', 'faith', 'form', 'true', 'edges', 'death', 'blank', 'ire', 'quot', 'de', 'work', 'calf', 'vol', 'heart', 'york', '8th', 'great', 'calf', 'street', 'red', 'age', 'notes', 'war', 'vo', 'rare', 'thou', 'li', 'mol', 'london', 'day', 'year', 'full', 'ppl', 'life', 'poems', 'work', 'the', 'poet', 'love', 'printed', 'man', 'god', 'series', 'faith', 'form', 'true', 'edges', 'death', 'blank', 'ire', 'quot', 'de', 'work', 'calf', 'vol', 'heart', 'york', '8th', 'great', 'calf', 'street', 'red', 'age', 'notes', 'war', 'vo', 'rare', 'thou', 'li', 'mol', 'london', 'day', 'year', 'full', 'ppl', 'life', 'poems', 'work', 'the', 'poet', 'love', 'printed', 'man', 'god', 'series', 'faith', 'form', 'true', 'edges', 'death', 'blank', 'ire', 'quot', 'de', 'work', 'calf', 'vol', 'heart', 'york', '8th', 'great', 'calf', 'street', 'red', 'age', 'notes', 'war', 'vo', 'rare', 'thou', 'li', 'mol', 'london', 'day', 'year', 'full', 'ppl', 'life', 'poems', 'work', 'the', 'poet', 'love', 'printed', 'man', 'god', 'series', 'faith', 'form', 'true', 'edges', 'death', 'blank', 'ire', 'quot', 'de', 'work', 'calf', 'vol', 'heart', 'york', '8th', 'great', 'calf', 'street', 'red', 'age', 'notes', 'war', 'vo', 'rare', 'thou', 'li', 'mol', 'london', 'day', 'year', 'full', 'ppl', 'life', 'poems', '

analytics on various unions and intersections of sets, as this description suggests, and has the potential to facilitate hands-on-discovery on the part of the instructor's students.

HTRC has been developed with digital humanists in mind, to overcome some of the technical, logistic, and accessibility hurdles present in large-scale text analysis. As it moves forward, the feedback of scholars is being listened to and solicited in order to meet the needs of scholars. However, the infrastructure that is currently in place already makes it a valuable tool for search.

References

1. **Milic, Louis Tonko** (1967). *A Quantitative Approach to the Style of Jonathan Swift*. Mouton: Walter de Gruyter.
2. **Burrows, John Frederick** (1987). *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method*. Oxford: Clarendon Press.
3. **Moretti, Franco** (2013). *Distant Reading*. London: Verso.
4. **Jockers, M.** (2013). *Macroanalysis: Digital Methods and Literary History*. University of Illinois Press.
5. **Belasco, Susan** (2011). *Whitman's Poems in Periodicals: Prospects for Periodicals Scholarship in the Digital Age*. in Earhart, Amy E. and Andrew Jewell. *The American Literature Scholar in the Digital Age*. Ann Arbor, MI: University of Michigan Press. p. 54.
6. **Michel, Jean-Baptiste et al.** (2011) *Quantitative Analysis of Culture Using Millions of Digitized Books*. Science, Vol. 331 No. 6014, 4 January 2011. pp. 176-182.
7. **Auvil, Loretta, Boris Capitanu, Matthew Jockers, Ted Underwood, and Ryan Heuser.** (2011) SEASR Analytics. Presentation at the Chicago Colloquium on Digital Humanities and Computer Science, Chicago, Illinois. November 19. chicagocolloquium.org/wp-content/uploads/2011/11/dhcs2011_submission_17.pdf . Retrieved: Oct 30, 2013.

Critical editing with TXSTEP

Ott, Wilhelm

Universität Tübingen, German

Ott, Tobias

Media University Stuttgart, Germany

In the "Afterword" to his 1984 edition of James Joyce's "Ulysses", Hans Walter Gabler gives a short outline of how he collected the variant readings contained in the different sources, how he used them for establishing, in two steps, the critical text, leaving most of the mechanical work (even the automatic insertion of the diacritic marks for the genetic variants) to the computer, and checking, by subsequent machine collation, the manual work which was carried out interactively at the computer console. According to Gabler, "the systematic and comprehensive reliance on computer aid ... has drastically reformed the editing process... Without it, this edition would neither be as accurate as we hope it is ... nor so rich in recorded facts" (p. 1909).

The TUSTEP tools Gabler had used more than 30 years ago have constantly been adapted in close collaboration with many editorial projects to their respective requirements and to changing technologies like PostScript and PDF for output or encoding standards like SGML, XML, TEI and Unicode. They have successfully been used for the preparation of many other critical editions; www.tustep.uni-tuebingen.de/ed3.html lists more than 750 volumes of printed editions published between 1972 and 2013 prepared and/or typeset with TUSTEP. They include works written in languages using non-latin alphabets, like greek (e.g. the 28th edition of Nestle-Aland, Novum Testamentum Graece, published in 2012), hebrew (e.g. the Mishna edition published by Michael Krupp) and arabic (Kitab al-Adad al-musamma..., ed. 2012 by Gunhild Graf). Current editorial projects relying on these tools include the works of Marx and Engels, the letters of Philipp Melanchthon, the works of Christoph Martin Wieland, of Albertus Magnus, the

philosophical works of Gottfried Wilhelm Leibniz and many others.

The TEI wiki judges the use of TUSTEP for the preparation of critical editions as follows:

- Advantage: does the job
- Drawback: very difficult to learn.

According to Willard McCarty (Humanities Computing 2005, p. 217), main reasons for these difficulties are the language of documentation and the complexity of the interface.

This mentioned drawback has in the meantime lost much of its impact:

At DH 2012, we presented the prototype of a modern XML-based interface to these tools, called TXSTEP. It both removes the language barrier and provides an user interface which an up-to-date established syntax. It allows the user to take advantage of the typical benefits of working with an XML editor, like content completion, highlighting, showing annotations, and verifying the code. The underlying XML schema contains extensive annotations and documentation on the purpose and syntax of the single functional elements available for building a TXSTEP script. When using a modern XML editor like oXygen, these annotations are shown automatically in a popup window while developing a TXSTEP script, so offering to a considerable degree a self teaching environment.

The poster session demonstrates how to use these tools for supporting the single steps required for the preparation of a critical edition:

- Collating witnesses / collecting variant readings
- Evaluating the collation results
- Constitution of edition text
- Compilation of apparatuses
- Preparation of indexes
- Preparation of printer's copy
- Publishing the text with apparatus(es) in print and/or for the web

As text basis for this demonstration, we chose a freely invented scenario: in order to have available a short example showing the whole spectrum of different types of variant readings, we copied a passage from vol. 4 of the edition of the works of Friedrich Schelling, a German philosopher, which had been typeset with TUSTEP in 1988, and labelled it as „version A“. In addition, we invented two other witnesses B and C for this same passage by copying it to separate files, there carrying out systematic replacements of single characters (so, the initial upper case Umlauts which in the 1988 edition have been written as Ae, Oe, Ue, have been converted to Ä, Ö, Ü) and other orthographic corrections (e.g., replacing th by t, y by i, c by k), changed the punctuation (inserting or deleting a comma etc.) and, in addition, made other modifications by inserting / omitting / replacing / transposing whole words or sentences. Starting from this material, the demo will show how the preparation and publication of a critical edition, be it in print or on the web, can profit from the power of TXSTEP – which provides powerful tools not only for editing but also for text analysis, for indexing, for lexicographic or bibliographic work and for publishing the results.

As a first step of the editorial work, we have to collect the variant readings by comparing the text of the sources, using the TXSTEP module COMPARE and specifying the details needed for our purpose (word-by-word collation; full references for the lemma location; abbreviate lemmata comprising more than 5 words; add 1 word of context for insertions; include a version-id for version B), as shown in fig. 1.

```

<?xml version="1.0" encoding="UTF-8"?>
<script xmlns="http://www.xstep.org"
  xsi:schemaLocation="http://www.xstep.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.xstep.org/2001/XMLSchema-instance"
  file:///C:/users/txstep/schema/txstep.xsd">
  <!-- Compare more than two (xml-)sources, output = diff-files with TEI
  inspired tags -->
  <variables>
    [8 lines]
  </variables>
  <compare versiona="$version-a" versionb="$version-b" variants="$diff-ab"
    type="readings" listing=".">
    <compare-units mode="word"/>
    <locations abbreviate="no" word-number="1">
      <add-context>
        <omission length-before="0" length="1" />
        <replacement length-before="0" length="1" />
        <insertion length-before="1" length="1" />
      </add-context>
      <pos-ind-b value="yes"/>
      <version-id>B</version-id>
    </compare>
    <display file="$diff-ab"/>
  <compare versiona="$version-a" versionb="$version-c" variants="$diff-ac"
    type="readings" listing=".">
    [10 lines]
  <collate source="$version-a" mode="single-witness"
    diff-file="$diff-ab'$diff-ac" listing="$diff-listing">
    [2 lines]
  <display file="$diff-listing" type="listing"/>
</script>

```

Fig. 1: COMPARE script

COMPARE compares two files only and produces two different kinds of output: a "comparison protocol" or listing which may be displayed on screen or printed on paper, showing the text of version A and below that the text of version B. Differences between the two versions are marked between the two lines. This listing (see fig. 2) is useful for checking the results of a correcting step or of automatic transformation of a text.

```

0.42   <u3>Allgemeine Uebersicht
  ==>   *
  ==>   <u3>Allgemeine Uebersicht
0.50   Gegners eben so  werth ist, als aus ihrem eignen, die bei
  ==>   - +*** - - ****
0.48   Gegners ebensoviel wert ist wie aus ihrem eignen, die bei
  ==>   - +*** - - ****
0.51   Untersuchungen jeder Art, - sie seyen groß oder klein, mehr
  ==>   - + * *
0.49   Untersuchungen jeder Art, - sie seien groß oder klein, mehr
  ==>   - + * *
0.54   verdammen, sobald ihnen bewiesen ist, daß sie  geirrt haben.
  ==>   ++++
0.52   verdammen, sobald ihnen bewiesen ist, daß sie sich geirrt haben.
  ==>   ++++
0.55   Er bekümmert sich nicht um kleine engherzige  Menschen, die
  ==>   -----
0.53   Er bekümmert sich nicht um  engherzige kleine Menschen, die
  ==>   -----
0.56   ihre Untersuchungen als eine aufgegebene Lection, oder als
  ==>   *
0.54   ihre Untersuchungen als eine aufgegebene Lektion, oder als
  ==>   *
0.57   ein Tagewerk betreiben, von dem sie nichts weiter als Lob
  ==>   *****
0.55   ein Tagewerk betreiben, von dem sie nichts weiter als Nahrung
  ==>   *****
0.58   oder Nahrung erwarten, die bei jeder Erweiterung des
  ==>   *****
0.56   oder Lob  erwarten, die bei jeder Erweiterung des
  ==>   *****
0.59   menschlichen Wissens nicht sowohl die Irrthümer, die sich so
  ==>   -
0.57   menschlichen Wissens nicht sowohl die  Irrtümer, die sich so
  ==>   -
0.65   thörlicht; jenes, weil es der Mühe nicht lohnt, dieses, weil
  ==>   -
0.63   töricht;  dieses, weil

```

Fig. 2: COMPARE protocol

For the preparation of critical editions, a different kind of output is more useful (see fig. 3): the differences found by comparison are written to a text file in a syntax which contains all the information necessary for further processing and merging into them the differences found by comparing the same version A to the text of other witnesses.

```

- <reading loc="0.50,2-0.50,3">
  <lem>eben so</lem>
  <rdg loc="0.48,2" mode="replace" wit="B">ebensoviel</rdg>
</reading>
- <reading loc="0.50,4">
  <lem>werth</lem>
  <rdg loc="0.48,3" mode="replace" wit="B">wert</rdg>
</reading>
- <reading loc="0.50,5">
  <lem>ist,</lem>
  <rdg loc="0.48,4" mode="replace" wit="B">ist</rdg>
</reading>
- <reading loc="0.50,6">
  <lem>als</lem>
  <rdg loc="0.48,5" mode="replace" wit="B">wie</rdg>
</reading>
+ <reading loc="0.51,3">
+ <reading loc="0.51,3">
- <reading loc="0.51,5">
  <lem>seyen</lem>
  <rdg loc="0.49,6" mode="replace" wit="B">seien</rdg>
</reading>
- <reading loc="0.54,7">
  <lem>sie</lem>
  <rdg loc="0.52,8" mode="insert" wit="B">sich</rdg>
</reading>
- <reading loc="0.55,6">
  <lem>kleine</lem>
  <rdg loc="0.53,6" mode="omit" wit="B"/>
</reading>
- <reading loc="0.55,7">
  <lem>engherzige</lem>
  <rdg loc="0.53,8" mode="insert" wit="B">kleine</rdg>
</reading>
- <reading loc="0.56,6">

```

Fig. 3: differences file with TEI compatible tags

A listing generated from the difference files produced by comparing the text of more than two witnesses is shown in fig. 4.

```

0.42   <u3>Allgemeine Uebersicht
  [B] ===== Übersicht
  [C] ===== Übersicht
0.43   <br/>der neuesten philosophischen Litteratur.</u3>
0.44   <absenk/>
0.45   <u2><kapitaelchen>Einleitung.</kapitaelchen></u2>
0.46   <abs>Der Verfasser, dem die Ausarbeitung dieses Artikels
0.47   Übertragen ist, kann sich über den Zweck desselben sehr kurz erklär
0.48   <abs>Er schreibt nur für diejenigen, die vor allen Dingen
0.49   <kursiv>Wahrheit</kursiv> wollen, denen sie aus dem Munde des
0.50   Gegners eben so  werth ist, als aus ihrem eignen, die bei
  [B] ===== ebensoviel wert ist wie ===== ===== =====
  [C] ===== ebensoviel ===== ist wie ===== ===== =====
0.51   Untersuchungen jeder Art, - sie seyen groß oder klein, mehr
  [B] ===== ===== ===== Art, - === seien ===== ===== =====
  [C] ===== ===== ===== Art, - === seien ===== ===== =====
0.52   oder minder wichtig - nicht ihr Individuum in Anschlag
0.53   bringen, und die immer die ersten sind, sich selbst zu
0.54   verdammen, sobald ihnen bewiesen ist, daß sie  geirrt haben.
  [B] ===== ===== ===== ===== ===== ===== sich ===== =====
  [C] ===== ===== ===== ===== ===== ===== sich ===== =====
0.55   Er bekümmert sich nicht um kleine engherzige  Menschen, die
  [B] ===== ===== ===== ===== ===== ===== kleine ===== =====
  [C] ===== ===== ===== ===== ===== ===== kleine ===== =====
0.56   ihre Untersuchungen als eine aufgegebene Lection, oder als
  [B] ===== ===== ===== ===== ===== ===== Lektion, ===== =====
  [C] ===== ===== ===== ===== ===== ===== Lektion, ===== =====
0.57   ein Tagewerk betreiben, von dem sie nichts weiter als Lob
  [B] ===== ===== ===== ===== ===== ===== Nahrung
  [C] ===== ===== ===== ===== ===== ===== Nahrung
0.58   oder Nahrung erwarten, die bei jeder Erweiterung des
  [B] ===== ===== ===== ===== ===== =====

```

Fig. 4: Listing showing differences of more than two witnesses

As the next step, shown in fig. 5, we try to classify the variants found. We try to differentiate between the various types of variants mentioned above when describing our text basis: we write variants concerning initial umlauts only to a separate file which we decide that its content will not be part of the apparatus, but will be handled in the preface. The merely orthographic variants and the variants concerning different punctuation only will get a respective attribute which allows to either also omit them from the apparatus, or to list them in a separate apparatus level. The remaining variants are those which should be listed in the main apparatus.

```

1 <?xml version="1.0" encoding="UTF-8"?>
2 <script xmlns="http://www.xstep.org"
3   xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
4   xsi:schemaLocation="http://www.xstep.org
5   file:///C:/users/txstep/schema/txstep.xsd">
6   <!-- distribute readings (resulting from txstep-script cmp3.xml) to different
7   files:
8     Umlaut spelling only, punctuation only, orthography only, "real" variants -->
9   <variables> [10 lines]
10  <transform source="$diffab" destination="$dest1$dest2"
11    save-params="$pars" option="erase">
12    <insert-at-start>{{variants}}</insert-at-start> <!-- for file $dest1 -->
13    <insert-at-start>{{umlauts}}</insert-at-start> <!-- for file $dest2 -->
14
15    <pass> <!-- preliminary pass: skip lines not starting with "reading">
16    <pass name="umlaut"> <!-- first pass: readings which differ in spelling of
17    upper case Umlauts only; output to file dest2--> [35 lines]
18    <pass name="punctuation"> <!-- second pass: readings which differ in
19    punctuation only; add attribute type="punctuation", output to dest1 --> [19 lines]
20    <pass name="ortho"> <!-- third pass: readings which differ from lemma
21    in th vs. t, c vs. k, y vs. i only; add attribute type="orthographic", output to
22    dest1; [32 lines]
23
24    <insert-at-end>{{/variants}}</insert-at-end>
25    <insert-at-end>{{/umlauts}}</insert-at-end>
26  </transform>

```

Fig. 5: classifying variant readings

The result of this step are two files: one containing the variants concerning (in our case) the initial umlauts only, and a second file containing the raw material to be used for building the apparatus entries. Fig. 5 shows the procedure for the differences between version A and B only (lines 17-122). There follows the same `<transform>` for the differences between version A and C, and a further module for sorting and merging the variants found in version B and version C in ascending order of lemma location, of type of variant, and of witness ID.

When we transform these sorted and merged records of variant readings automatically, without philological inspection and revision, into apparatus entries for a printed critical edition and insert them into our edition text (for which, in this demo, the unaltered text of version A will serve), we get a file of which a detail is shown in fig. 6.

```

- <u3>
  Allgemeine Uebersicht
  <br/>
  der neuesten philosophischen Litteratur.
</u3>
<u2>Einleitung.</u2>
<abs>Der Verfasser, dem die Ausarbeitung dieses
  Artikels übertragen ist, kann sich über den Zweck
  derselben sehr kurz erklären.</abs>
- <abs>
  Er schreibt nur für diejenigen, die vor allen Dingen
  <kursiv>Wahrheit</kursiv>
  wollen, denen sie aus dem Munde des Gegners eben
  - <app mode="long">
    <lem>eben so</lem>
    <rdg mode="replace" wit="B C">ebensoviel</rdg>
  </app>
  so
  <anchor xml:id="B C"/>
  werth
  - <app>
    <lem>werth</lem>
    <rdg mode="replace" wit="B"
      type="orthographic">wert</rdg>
  </app>
  ist,
  - <app>
    <lem>ist,</lem>
    <rdg mode="replace" wit="B C"
      type="punctuation">ist</rdg>
  </app>
  als
  - <app>
    <lem>als</lem>
    <rdg mode="replace" wit="B C">wie</rdg>
  </app>
  aus ihrem eignen, die bei Untersuchungen jeder Art, -
  - <app>

```

Fig. 6: edition text (version A) with inserted apparatus entries

With a further script (not shown here), making use of the powerful typesetting engine of TUSTEP, this file is transformed to a PostScript or PDF file (see fig. 7), showing at page bottom the apparatus entries linked to the text by means of line numbers printed in the margin of the edition text. As stated above, in this example, we omitted only the variants concerning the writing of the initial upper case Umlaut only; for the other variants, three apparatus levels have been provided, the first one showing the variant readings which may affect the meaning or interpretation of the text; merely orthographic variants are listed in the second apparatus, the third apparatus contains variants regarding punctuation only.

Allgemeine Uebersicht
der neuesten philosophischen Litteratur.

EINLEITUNG.

Der Verfasser, dem die Ausarbeitung dieses Artikels übertragen ist, kann sich über den Zweck desselben sehr kurz erklären.

Er schreibt nur für diejenigen, die vor allen Dingen *Wahrheit* wollen, denen sie aus dem Munde des Gegners eben so werth ist, als aus ihrem eignen, die bei Untersuchungen jeder Art, — sie seyen groß oder klein, mehr oder minder wichtig — nicht ihr Individuum in Anschlag bringen, und die immer die ersten sind, sich selbst zu verdammen, sobald ihnen bewiesen ist, daß sie geirrt haben. Er bekümmert sich nicht um kleine engherzige Menschen, die ihre Untersuchungen als eine aufgegebene *Lection*, oder als ein Tagewerk betreiben, von dem sie nichts weiter

8 eben so] ebensoviel *B C* 9 als] wie *B C* Art, —] Art, *B C add. — B C* 13 sie] add. sich *B C* kleine] om. *B C* 13–14 engherzige] add. kleine *B C*

8 werth] wert *B* 10 seyen] seien *B C* 15 Lection,] Lektion, *B C*

9 ist,] ist *B C*

3

Fig. 7: typeset Edition with 3 apparatus levels at page bottom

Of course, without investing further philological effort, the results are less than satisfying, as shown in fig. 7 for lines 13–14: the text has "kleine engherzige", versions B and C each showing "engherzige kleine". The apparatus says that, in line 13, version B and C omit "kleine" and insert, after "engherzige" ending in line 14, the word "kleine". The reason is that the word-by-word comparison has produced this result. In the apparatus, the two entries should however be transformed into a single one, showing an inversion of the two words "kleine engherzige" to "engherzige kleine".

This means: the procedures shown in the demo can only make available a reliable material basis for the philological work indispensable for responsible critical work on an edition. At the same time, the scope of work of these procedures can in every detail be specified according to the needs of an actual editorial project, thus saving time for manual work and at the same time providing a reliable material basis for controlled work.

Fig. 8 shows an example of the same text presented as an online edition (also generated automatically, without critical intervention) in html form. In the left column, we have the edition text (version A in our case), highlighting the words where one of the other versions (B and C in our case) show differences. A click at one of the highlighted words makes visible the respective location in the apparatus frame below the text, where a click at the witness code opens, in the right hand frame, the text of the selected version and shows it from the location containing the respective variant reading.

Version A	Version B
Allgemeine Uebersicht der neuesten philosophischen Litteratur.	Allgemeine Übersicht der neuesten philosophischen Litteratur.
Einleitung.	Einleitung.
Der Verfasser, dem die Ausarbeitung dieses Artikels übertragen ist, kann sich über den Zweck desselben sehr kurz erklären.	Der Verfasser, dem die Ausarbeitung dieses Artikels übertragen ist, kann sich über den Zweck desselben sehr kurz erklären.
Er schreibt nur für diejenigen, die vor allen Dingen <i>Wahrheit</i> wollen, denen sie aus dem Munde des Gegners <u>eben so</u> werth ist, als aus ihrem eignen, die bei Untersuchungen jeder <u>Art</u> , — sie <u>seien</u> groß oder klein, mehr oder minder wichtig - nicht ihr Individuum in Anschlag bringen, und die immer die ersten sind, sich selbst zu verdammen, sobald ihnen bewiesen ist, daß sie geirrt haben. Er bekümmert sich nicht um <u>kleine engherzige</u> Menschen, die ihre Untersuchungen als eine aufgegebene <u>Lection</u> , oder als ein Tagewerk betreiben, von dem sie nichts weiter	Er schreibt nur für diejenigen, die vor allen Dingen <i>Wahrheit</i> wollen, denen sie aus dem Munde des Gegners <u>ebensoviel</u> werth ist wie aus ihrem eignen, die bei Untersuchungen jeder <u>Art</u> , — sie <u>seien</u> groß oder klein, mehr oder minder wichtig - nicht ihr Individuum in Anschlag bringen, und die immer die ersten sind, sich selbst zu verdammen, sobald ihnen bewiesen ist, daß sie <u>sich</u> geirrt haben. Er bekümmert sich nicht um <u>engherzige kleine</u> Menschen, die ihre Untersuchungen als eine aufgegebene <u>Lektion</u> , oder als ein Tagewerk betreiben, von dem sie nichts weiter
Varianten zu Version A	Varianten zu Version B
eben so] ebensoviel (b) (c)	ebensoviel] eben so (a)
werth] wert (b) (c)	wert] werth (a)

Fig. 8: edition in html frames

Also the right-hand frame, in this case presenting version B, shows an apparatus frame containing however the differences to version A only and not to the other witnesses. Also here, the above mentioned inversion "kleine engherzige" vs. "engherzige kleine" is not marked as an inversion: "um" - though identical to the text in version A - is highlighted because version A has inserted "kleine" after "um", and "kleine" is highlighted because it is missing after "engherzige" in version A.

Of course, editorial work cannot rely on the steps only which are shown in this demo. Alphabetic lists of word forms occurring in the single witnesses, e.g., or sorting the lemma-variant-pairs alphabetically could help to reveal e.g. spelling variants typical for certain geographical regions where a witness has been transcribed and e.g. help to build groups of witnesses for texts with a great number of witnesses. Also for these and similar tasks, TXSTEP provides not ready-made solutions but a set of tools for relatively elementary steps of text data processing – tools whose scope of work can be specified in detail according to the needs of each step, and which can be combined in (almost) arbitrary ways to provide solutions also for complex tasks. A list of the basic modules is shown in fig. 9, showing the popup window if you put the cursor in the root tag of a TXSTEP script (in this case, I did this using the script shown in fig. 1).

The screenshot shows a Windows-style application window titled "cmp3.xml [C:\Users\stephe\scripts\cmp3.xml] - <@Xygen> XML Editor (Aussichtlich akademische Nutzung)". The main pane displays an XML document with numerous comments explaining various modules provided by TXSTEP. These modules include:

- txstep**: Provides a XML-based interface for TUSTEP, the Tübingen System of Textprocessing tools. It includes modules for:
 - COMPARE**: Compare two text versions, list the differences, file them for further processing.
 - COLLATE**: Lists the differences between more than one versions of the same text, found by previous pairwise COMPARE operations, displaying them in a synopsis of the base text and the readings of the other versions.
 - CORRECT**: Correct texts in batch mode by means of previously filed correcting instructions.
 - PREPARE-INDEX**: Break down text into sortable index entries, or extract text parts marked as index entries and prepare them for sorting.
 - PREPARE-SORT**: Text units, consisting of one or more input records, are prepared for sorting.
 - SORT**: Sort files containing records prepared by PREPARE-SORT or PREPARE-INDEX.
 - GENERATE-INDEX**: Generate indexes from sorted index entries.
 - TRANSFORM**: Select records or groups of records or text units consisting of more than one record from input file, rearrange text parts, replace character strings, write output to one or more files.
 - INSERT**: Integrate text parts, which are identified by acronyms, from a second file into the text of the source file by replacing or expanding the respective acronyms.
 - RENUMBER**: Update (running) numbers and the respective references contained in a text; update references to page-line-numbers after the page-line-division has been changed.
- Original Tustep and Tuscript scripts can be integrated.

Fig. 9: showing, as popup, the modules provided by TXSTEP

Both TXSTEP and TUSTEP, the TUEbingen System of TEx Processing tools, are open source under the Revised BSD License and can be downloaded from the TUSTEP homepage[6]. The TXSTEP installation package contains in addition a set of 80 exercises, covering tasks like file transformation and extraction of information, collation of different versions of the same text, evaluation of collation results, index generation and sorting.

References

- <http://www.bbaw.de/forschung/mega/>
- <http://www.haw.uni-heidelberg.de/forschung/forschungsstellen/melanchthon/edition.de.html>
- <http://www.wieland-edition.uni-jena.de/>
- http://institut.albertus-magnus-web.de/643_0/editiocoloniensis.html
- <http://www.uni-muenster.de/Leibniz/seite2.html>
- www.tustep.org

Visualization As a Bridge to Close Reading: The Audience in The Castle of Perseverance

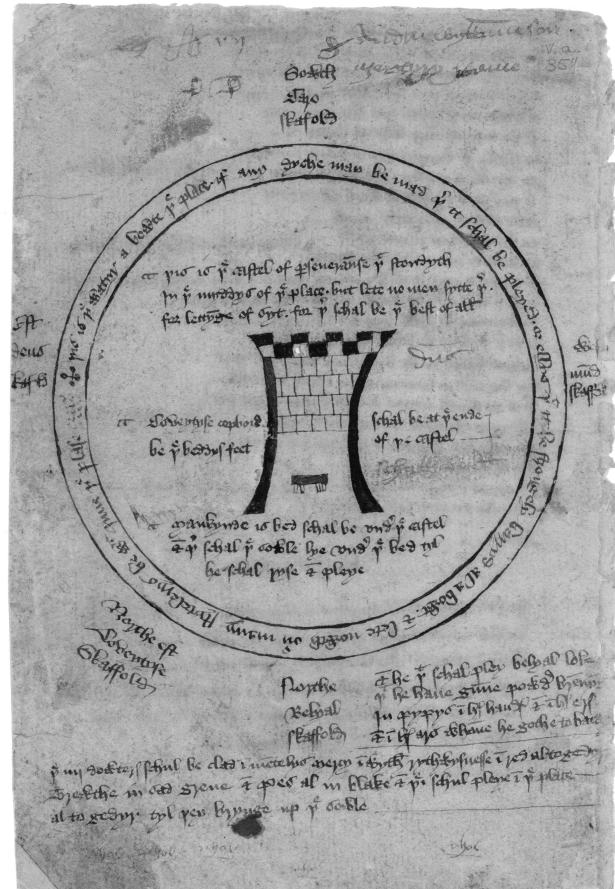
Peterson, Noah
noah.peterson@tamu.edu
 Texas A&M University

Visualization As a Bridge to Close Reading: The Audience in *The Castle of Perseverance*

1. Introduction

In thinking of visualizations and how they can help bridge the gap between traditional close readings of texts and digital projects, Matthew Kirschenbaum asks, "What patterns would be of interest to literary scholars? ... How would we evaluate the effectiveness of our visualizations, or the software in general? Is it succeeding if it surprises us with its results, or if it doesn't?"¹

Many scholars have noted the importance of approaching a visualization project with an appropriate amount of scholarly attention, and it is my intention to show how the activity of creating a visualization of a text and the visualization itself can highlight areas of the text which would benefit from a traditional close reading.^{2 3 4} The text I have chosen for this visualization-prompted close reading is *The Castle of Perseverance*, a 15th century morality play which takes the form of a *locus-and-platea* play. Creating a visualization of the network of characters within the play leads us to inquire into the role of the audience of the play, both the original audience of medieval spectators and contemporary readers or watches of the play.



The Castle of Perseverance stage plan, Macro MS folio 191v.

2. Tool

The tool I have chosen for the visualization is Gephi, an open-source graphic visualization software. Gephi is "an interactive visualization and exploration platform for all kinds of networks and complex systems, dynamic and hierarchical graphs."⁵ I have two main reasons for this choice. The first is that Gephi is a relatively easy program to work with, while also being quite powerful and allowing a nuanced approach to the visualization process. The second reason for this choice is the set of algorithms which come with Gephi, two of which I will be using in my visualizations. One measures "Betweenness Centrality," a measure of the size of the network and the average path length between nodes, in this case characters, in the network. The second algorithm measures modularity and identifies community groups within the network. Different critical approaches to visualizing will affect the ways in which these two algorithms interpret the data and the resulting visualization will lead to different conclusions.

3. Theoretical Issues

One of the benefits of creating a visualization of the network within a play is that "nothing ever disappears. What is done, cannot be undone ... The past becomes the past, yes, but it never disappears from our perception of the plot."⁶ In the case of *The Castle of Perseverance*, however, this permanence of action that exists in a visualization can cause some problems.

Medieval morality plays, by nature, are heavily didactic and often address the audience directly. In *The Castle of Perseverance*, over 800 lines, more than one-fifth of the play, are directed at the audience. Throughout the play, the audience is addressed both directly and indirectly, "all the men that in this wold wold thryve" (521), "Lordyngys" (1425), and "all men" (3694) to give but a few examples.⁷ Should a visualization of the network of characters in a play include the audience of that play? They have no speaking role and would certainly not be included in any list of *dramatis personae*. Leaving the audience out of the network leads to the network which appears in figure 2.

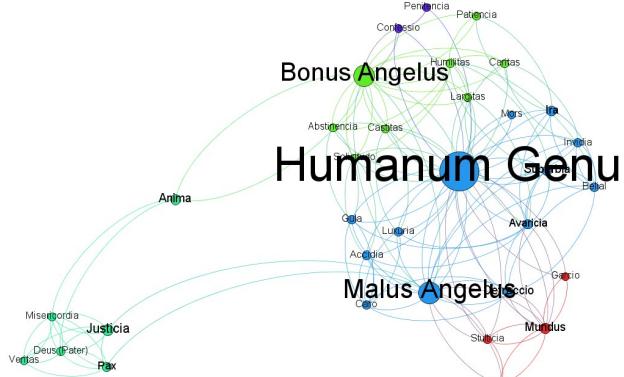


Fig. 1: The Castle of Perseverance network, without the audience as a node.

If we are troubled by leaving out such a large portion of the play and create a network which includes the audience, the result is figure 3.

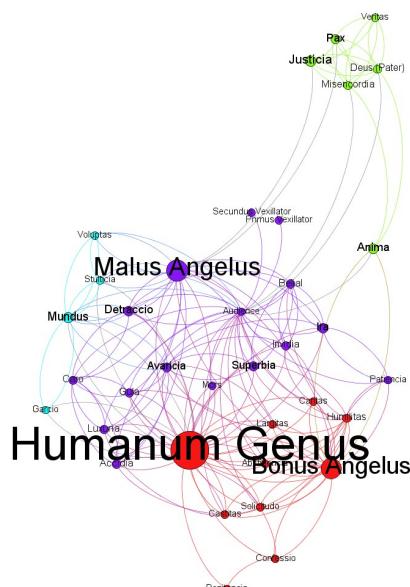


Fig. 2: The Castle of Perseverance network with the audience included as a node.

The interesting difference between these visualizations is the members of the various communities as identified by Gephi. In figure 2, Humanum Genus, the mankind figure, is grouped with the bad angle and most of the seven deadly sins. In figure 3, however, Humanum Genus is grouped with the good angle and the seven virtues; this is also how the play ends, with Humanum Genus asking for God's mercy and ascending to heaven. The audience in some way, according to the communities identified by Gephi, is a factor in Humanum Genus ending the play in God's grace. These two maps of the network within *The Castle of Perseverance* cause us to turn back to the text for a close reading of the ways in which the audience appears as a character through the efforts of the dramatic characters of the play.

References

1. Kirschenbaum, M. *Poetry, Patterns, and Provocation: The Nora Project*. The Valve. January 12, 2006. Web. October 31, 2013.
2. Jessop, Martyn (2008). *Digital Visualization As a Scholarly Activity*. Literary and Linguistic Computing 23.3: 281-193. Print.
3. Rieder, Bernhard and Theo Rohle (2012). *Digital Methods: Five Challenges*. Understanding Digital Humanities. Ed. David M. Berry. New York: Palgrave Macmillan. Print. 67-84.
4. Scully, D. and Bradley M. Pasanek (2008). *Meaning and Mining: The Implicit Assumptions in Data Mining for the Humanities*. Literary and Linguistics Computing 23.4: 409-424. Print.
5. www.gephi.org
6. Moretti, Franco. (2011). *Network Theory, Plot Analysis*. Stanford, Stanford Literary Lab.
7. Eccles, Mark (1969). *The Macro Plays*. London: Oxford University Press. Print.

How we work: a critical approach to program development to serve library/dh partnerships

Potvin, Sarah

spotvin@library.tamu.edu

Texas A&M University, United States of America

Herbert, Bruce

Texas A&M University, United States of America

Earhart, Amy

Texas A&M University, United States of America

How we work: a critical approach to program development to serve library/dh partnerships

"Some people say, "what you're developing is a Sciences-based model" and no, actually [what] we're doing is this experiential based model that is very much geared toward working in the Humanities. It's just different kind of pedagogy, a different kind of learning, a different kind of research that we're engaged in."

-Quoted in Siemens, Cunningham, Duff, Warwick (2011).

In this short paper, we will analyze studies of how digital humanists and scientists work, testing the oft-referenced distinctions and similarities claimed between science and dh models. The three authors of the paper -- a librarian, scientist, and digital humanist -- bring expertise in different domains and traditions to bear. While much dh work has focused on the expertise of technologists and computer scientists, less attention has been paid to the ways that scientific disciplinary impacts digital humanities work. And yet science and dh exert influences on one another, particularly as practices and tools developed in the sciences are imagined, borrowed, and manipulated by dh, but also as practices and insights from the humanities are applied to science inquiry. We resist a model wherein science is imported into the humanities unidirectionally, without reciprocal influence.

The impact of partnerships between university-based dh and libraries are strengthened through an understanding of these influences on scholarly practices in dh. As Palmer and Cragin (2008, p. 165) argue, "Understanding the nature of information practices and their relation to the production of scholarship is important for both theoretical and applied work in library and information science (LIS). Research on scholarly practices provides a foundation for the development of information systems, services, and tools to support scholarship and science ..." The applied goal of our research is the framing of the question of library-dh partnerships and the crafting of an approach to program development by the Texas A&M University Libraries that emphasizes a broad coalition of interested dh participants. We argue that examining program design through the theoretical lens of communities of practice will further enhance library and digital humanities partnerships.

A rich, existing literature considers co-authorship patterns and collaborative traditions in the digital humanities (Spiro, 2009; Nowviskie, 2011; Nyhan and Duke-William, 2013) and investigates the sociology of work for different disciplines. Recent scholarship has explored the idea of laboratories, an often cited trope in digital humanities literature (Earhart, n.d.). Additionally, epistemological models from the sciences and the humanities (Duschl and Grandy, 2008) provide insight into how disciplines frame and approach knowledge creation. The concept of data and their deployment in scientific and humanistic settings has been something of a touchstone for disciplinary bounding. Indeed, as Drucker argues, the humanities should reconceptualize data as "capta": "From this distinction, a world of differences arises. Humanistic inquiry acknowledges the situated, partial, and constitutive character of knowledge production, the recognition that knowledge is constructed, taken, not simply given as a natural representation of a pre-existing fact" (2011).

In our research, we will critically evaluate the comparisons drawn between epistemological and labor models in dh and the sciences. This approach is in keeping with recent work cautioning against importing from the sciences without an awareness of the complex systems behind them. Earhart cautions: "it is important not to romanticize the lab. ... while we might look to the laboratory as a model, we need to be critical about its implementation in our field" (n.d.). As Unsworth has written: "In humanities, we often emulate what we think the sciences do, but our emulation may not actually bear that much resemblance to the reality of what goes on in science. Often science looks more collaborative because a lot of people get together to write a grant proposal, but that

does not mean that they have necessarily figured out how to work together" (Unsworth and Tupman, 2012, p. 232. Cited in Earhart, 2014).

While the "imagined sciences" have not been a major focus on the dh literature, valuable existing work considers the influence of science under the guise of computer science, technology, and library sciences. A growing literature examines how librarians and humanists work on digital projects (Siemens, Cunningham, Duff, Warwick, 2011) and surveys proliferating models for dh-library and dh-librarian collaboration and programming, as well as partnerships and efforts to "reskill" around dh (Posner, 2013; Bryson, Posner, St. Pierre, Varner, 2011). In addition to providing data on influences, these sources will further ground our program design.

Application

University initiatives experience a high rate of failure (Kezar and Eckel, 2002), designed as they often are with simplistic or inaccurate change models. In developing our program, we aim to meet the imperative to design with a community's needs and practices as its focus, taking the factors and interactions that will affect the success of the program into consideration.

To that end, we will need to match our programming to the particular context of Texas A&M University. As a first step, we will attempt to serve this need through the creation of an institutional profile assessing publication patterns of university faculty. How do rates of, for example, co-authorship or patterns of citations on blogs for humanities faculty at A&M compare with humanities faculty at other institutions? A land-grant, sea-grant, space-grant university with more than fifty thousand enrolled students, A&M has evolved from its historical emphasis on agricultural and mechanical education and become known as a science- and engineering-oriented research institution. Have expectations around the sciences, dominant as they are at the university level, shaped dh work at A&M more profoundly? Is this evidenced in publications?

Taken in combination with our creation of an institution profile, our research into the influences between the sciences and the humanities will provide a framework for the design of a library-dh program. Understanding the "science-based model" and its effects on dh will further inform the development of a system of personas matched to dh needs, barriers, and opportunities on campus. We believe that this research-based approach to dh partnership development will interest the larger dh community.

References

- Bryson, Tim; Posner, Miriam, St. Pierre, Alain; Varner, Stewart. (2011). *SPEC Kit 326: Digital Humanities*. (Association of Research Libraries).
- Drucker, Johanna. (2011). *Humanistic approaches to graphical display*. *Digital Humanities Quarterly* 5, no. 1.
- Duschl, R.A., and R.E. Grandy. (2008). *Teaching scientific inquiry: recommendations for research and implementation*. Rotterdam, The Netherlands: Sense Publishers
- Earhart, Amy. (n.d.) *The Digital Humanities as a Laboratory*. In *Humanities and the Digital*. Ed. David Theo Goldberg and Patrik Svensson. Under advance contract, MIT Press.
- Huang, Mu-Hsuan and Yu-Wei Chang. (2013). *Quantifying the value of knowledge exports from librarianship and information science research*. *Journal of Information Science* 39: 141-150.
- Kezar, A., & Eckel, P. (2002). *The effect of institutional culture on change strategies in higher education: Universal principles or culturally responsive concepts?* *The Journal of Higher Education*, 73(4), 435-460.
- Nowviskie, Bethany. (2012). *Evaluating Collaborative Digital Scholarship (or, Where Credit is Due)*. *Journal of Digital Humanities* 1, no. 4.
- Nyhan, Julianne and Oliver Duke-William (2013). *Joint and multi-authored publication patterns in the Digital Humanities*. *Digital Humanities 2013 Conference Proceedings*. dh2013.unl.edu/abstracts/ab-338.html

- Palmer, Carole L. and Cragin, Melissa H.** (2008). *Scholarship and disciplinary practices*. Annual Review of Information Science and Technology 42.
- Posner, Miriam.** (2013). *No Half Measures: Overcoming Common Challenges to Doing Digital Humanities in the Library*. Originally published in Journal of Library Administration 53, no. 1. Special issue: Digital Humanities in Libraries: New Models for Scholarly Engagement.
- Siemens, Lynne; Cunningham, Richard; Duff, Wendy; Warwick, Claire.** (2011) "A tale of two cities: implications of the similarities and differences in collaborative approaches within the digital libraries and digital humanities communities." Special issue: Papers from Digital Humanities 2010, King's College, London. LLC 26, no. 3: 335-348.
- Siemens, Lynne.** (2009). "It's a Team If You Use "Reply All": An Exploration of Research Teams in Digital Humanities Environments." LLC 24, no. 2: 225 -233.
- Sin, Joanna Sei-Ching.** (2011). *International coauthorship and citation impact: A bibliometric study of six LIS journals, 1980-2008*. Journal of the American Society for Information Science and Technology 62, no. 9: 1770-1783.
- Spiro, Lisa** (2012). "*This Is Why We Fight: Defining the Values of the Digital Humanities.*" In Matthew K. Gold, ed., Debates in the Digital Humanities (University of Minnesota Press).
- Sula, Chris Alen and Matt Miller** (2013). *Citation studies in the humanities*. Digital Humanities 2013 Conference Proceedings. dh2013.unl.edu/abstracts/ab-353.html
- Unsworth, John and Tupman, Charlotte.** (2012). Interview with John Unsworth, April 2011, carried out and transcribed by Charlotte Tupman. In Collaborative Research in the Digital Humanities, ed. Marilyn Deegan and Willard McCarty. London: Ashgate. pp. 231-9.

Seeing Dialogue: Network Visualization of Dramatic Texts

Powell, Daniel
djpowell@uvic.ca
University of Victoria

>Contexts for Humanities Visualization

Literary scholars are increasingly turning to graphical display of humanistic information as a way to encounter texts in new and engaging ways. Digitally facilitated humanities visualization presents literary critics with opportunities for new insight¹, while also opening practitioners to charges of wholesale importation of simplistic scientific methodologies.² This poster session outlines the rationale, method, and significance of a focused humanities visualization project in order to demonstrate how new techniques of visualization may be undertaken with literary texts to produce new and speculative "aesthetic provocations" in literary studies.³

Materials and Scope

In order to model and theorize how such a project might develop, I am producing a network visualization of the sixteenth-century play Ralph Roister Doister by Nicholas Udall. Using the open-source "interactive visualization and exploration" platform Gephi, I map the relationships between characters in the play based upon dialogue. In other words, this project structurally maps dialogue between characters, producing a network visualization that productively reconfigures the play to provoke new analytical responses. Following the example of Franco Moretti's work with Hamlet, these models prompt new insight into the "deep structures" of literary works. For Moretti, however, "the most important thing of all" about these reconfigurations is that they can be manipulated:

"one can intervene on a model; make experiments."⁴ [116, italics in original]. With a baseline visualization of dialogue established, I, following Jerome McGann, selectively intervene and "deform" the re-modeled text to continually reconfigure it. Such processes bring us "to a critical position in which we can imagine things about the texts that we didn't and perhaps couldn't otherwise know."⁵ [116] Motivated by Drucker and Nowviskie's call to "engage computing to produce new aesthetic provocations," I use Roister Doister to understand how humanities visualization may reconfigure our approach to literary inquiry.⁶

Methods of Production

In order to map the dialogic relationships between these characters in Gephi, I have created each character as a node in the network; these nodes are connected by directional dialogue originating at a particular node and terminating at another. Thus, the main characters Roister Doister and Merygreeke are nodes 1.0 and 2.0, for example. Within Gephi's data manipulation environment, a line of dialogue from the former to the latter would be mapped visually as a line [or an "edge"] from node 1.0 to node 2.0. Each exchange is recorded as tabbed data in Gephi, which is then used to produce visualizations. Visualizations can be produced for any discrete unit of the text, including scene, act, or the entirety of the play. This project produces visualizations of each of these divisions for comparative purposes. Following Moretti and McGann, I have undertaken selective deformations by, for example, removing various characters at different points, thereby revealing their network centrality. Criticism of Roister Doister has, for the most part, focused heavily on the play's dramaturgical debt to the classical comedies of Terence and Plautus, the extent to which those formal structures were successfully integrated with "native" elements of English drama, and the play's debts to the miles gloriosus ["braggart-soldier"] tradition of Classical comedy.⁷ ⁸ This project is in part an attempt to revitalize an ossified critical conversation as an example of how new techniques can vigorously re-engage old texts.

Significance of Anticipated Visualizations

Experimentation in visualization of textual works is a timely one. As can be seen from the growing use of the *Voyant* suite⁹ of analysis and visualization tools as well as the popular *Mapping the Republic of Letters* project,¹⁰ humanities visualization is a growing area of scholarly concern. Elijah Meeks has argued that "the shift from creating, annotating and analyzing archives to modeling systems can have a profound impact beyond the [admittedly high value of] usability of scholarly material developed during a digital humanities project."¹¹ [italics mine]. When humanities scholars have reached a certain point of visual literacy, we will begin to engage with such models in profoundly important ways. Indeed, these models may "provide a much more nuanced form of knowledge transmission than the raw datasets or interactive and dynamic applications typically presented as the future of digital scholarly media."¹² This project is an effort to explore how these new forms of knowledge transmission and analysis might impact literary inquiry.

References

1. **Manovich, Lev.** (2010). *What Is Visualization?* Poetess Archive Journal 2(1), n. pag.
2. **Drucker, Johanna.** (2011). *Humanities Approaches to Graphical Display*. Digital Humanities Quarterly 5(1), n. page.
3. **Drucker, Johanna, and Bethany Nowviskie.** (2004). *Speculative Computing: Aesthetic Provocations in Humanities Computing*. Companion to Digital Humanities. In Schreibman, Susan, Siemens, Ray, & Unsworth, John (eds). Oxford, Blackwell. Accessed 03 April 2012. Available at [journals.tdl.org/paj/index.php/paj/article/view/19/58]

- 4.5. **Moretti, Franco.** (2011). *Network Theory, Plot Analysis*. Stanford, Stanford Literary Lab.
6. **Drucker, Johanna, and Bethany Nowviskie.** (2004). *Speculative Computing: Aesthetic Provocations in Humanities Computing*. Companion to Digital Humanities. In Schreibman, Susan, Siemens, Ray, & Unsworth, John (eds). Oxford, Blackwell. Accessed 03 April 2012. Available at [journals.tdl.org/paj/index.php/paj/article/view/19/58]
7. **Boas, Frederick S.** (1933). *An Introduction to Tudor Drama*. Oxford, Clarendon.
8. **Brooke, C. F. Tucker.** (1911). *The Tudor Drama: A History of English National Drama to the Retirement of Shakespeare*. Boston, Riverside.
9. **Voyant Tools.** (2012). [voyant-tools.org]. Accessed October 2012.
10. *Mapping the Republic of Letters: Navigating Big Data from the Early Modern Period.* (2012). Available at [republicofletters.stanford.edu].
11. **Meeks, Elijah.** (2011). *More Networks in the Humanities or Did Books Have DNA?* Digital Humanities Specialist. Published 6 December 2011. Accessed 17 April 2012.
12. **Meeks, Elijah.** (2011). *More Networks in the Humanities or Did Books Have DNA?* Digital Humanities Specialist. Published 6 December 2011. Accessed 17 April 2012.

Discovering Old Maps Online and Transforming Them Into Digital Humanities Resources

Pridal, Petr

Klokán Technologies GmbH, Switzerland

Hundreds of thousands of historical maps have now been scanned and made available on-line by libraries and archives around the world, and this has been a great boon to anyone interested in the history of cartography. Despite this fact it is hard to find scanned maps covering area of interest in the large number of online catalogs, library systems and web presentations on the web. The traditional full text search engines, such as Google, are failing to index the scanned maps properly.

Old Maps Online (www.oldmapsonline.org) is a search system tailored just for historical maps. Pick a location on a world map, or type in a place-name, narrow the search by selecting a date range. A listing of all possible maps covering that location appears, ordered by best geographical match. Select a map, click on the link and you go directly to view the map on the original library's website.

You don't need to know who holds the map, just when and where in the world you want to look at. This system is designed to complement rather than compete with libraries' own search interfaces. The system is powered by the enhanced version of the MapRank Search technology and indexes over 130.000 scanned high-resolution maps already and this number grows.

Many major collections in the US, UK and elsewhere have agreed to contribute: The British Library, Harvard Library, National Library of Scotland, David Rumsey Map Collection, Dutch National Archives, Moravian Library, New York Public Library, Norman B. Leventhal Map Center at the Boston Public Library, National Library of Australia, etc. Cooperation with the Europeana project has started as well. Our aim is to include as many collections as possible, so map libraries and collectors are encouraged to participate.

To be able to index the scanned maps geographically, the minimal metadata (title, creator/publisher, date, identifier, and a stable url) plus geographic coordinates for the area covered must be known for each map.

We develop and maintain set of online tools targeting librarians, scholars and volunteers allowing to create the coordinates and assign precise location to the existing online zoomable maps. One of the tools is the Georeferencer online service, which allows rapid collaborative georeferencing, 3D visualization, annotation and accuracy analysis of scanned

online maps directly in a web browser environment, without the need to install any additional software on a local computer. The online visitors can help with the metadata enrichment and georeferencing of the scanned maps - and they are motivated with competitions, rewarding, community participation and recognition during this crowdsourcing effort.

With the presented online tools the enrichment and reuse of scanned maps is very straightforward. Recently the system has been improved and extended also with the functionality for favouriting the maps, for creating custom virtual map collections and for overlaying and comparing of the old maps. New tools for annotating and for exporting produced geodata or reusing the maps in a form of the standardized online services OGC WMS / OGC WMTS are integrated as well.

The Georeferencer service is applied in several institutions such as the British Library (London), the David Rumsey Map Collection (USA), the Moravian Library (Brno), the Nationaal Archief (The Hague), the National Library of Scotland (Edinburgh), and the Institut Cartografic de Catalunya (Barcelona).

Next to the service for discovery and indexing of these beautiful valuable resources, and next to the mentioned online tools for collaborative metadata enrichment of the scanned maps, this paper shows also practical applications of free and open-source software projects for online publishing and presentation of high resolution maps and vast collections of large raster images in general. Thanks to the modern web technologies such as HTML5 and WebGL it is now technically possible to bring the maps into web mashups and custom online applications directly.

The maps are undoubtedly extremely important documents for scholars, experts as well as interested general public. Once enriched with the precise geolocation and additional annotations, they are turned into a very practical reusable online resource. Applying presented available tools and methods in a large scale opens opportunities for completely new forms of research in the area of digital humanities, which has never been possible before.

Geographies of Access: Mapping the Online Attention to Digital Humanities Articles in Academic Journals

Priego, Ernesto

City University London, United Kingdom

Havemann, Leo

Birkbeck College, University of London, United Kingdom

Atenas, Javiera

SOAS, University of London, United Kingdom

Open access refers to the free access to and reuse of scholarly works. Peter Suber, who was the principal drafter of the Budapest Open Access Initiative (February 2002), and authored the book titled Open Access (2012), defines it as academic literature that is "digital, online, free of charge, and free of most copyright and licensing restrictions."

What proportion of peer-reviewed digital humanities research is published in open access journals? What proportion of digital humanities monographs and edited collections are available in electronic formats, and which, if any, are available in open access form? How do open access articles about the digital humanities compare in terms of citations/downloads to their toll-access counterparts? How, when, where and why do digital humanities scholars and the public engage in online attention to online academic articles about digital humanities and why does it matter? What kind of licensing and copyright agreements are digital humanities scholars subscribing to, and of those which ones allow and encourage open collaboration and reuse,

including text and data mining? What is the role of blogging in the digital humanities publishing landscape?

These are the questions guiding the research project whose findings we will visualise through an infographic. It will show the findings of a comparative, quantitative bibliometric analysis of a data set of academic articles about the digital humanities published between 2010 and 2013. The infographic will visualise the conclusions of an ongoing collaborative research project whose aims are to employ journal and article-level quantitative and qualitative analysis to determine whether alt-metrics can provide a holistic image of impact on diverse audiences.

The poster will also include a visualisation of the geographical distribution of online attention to the articles published on both journals, as well as other quantitative and qualitative data. The main objective of the poster will be to provide demographic data of online activity reflecting the attention paid to digital humanities research by other researchers, the media and the general public, providing much-needed data about where academic articles on digital humanities are published, which are the business models the chosen platforms have (toll-access, open access) as well as other information as presence or absence of digital identifiers, secure archiving, etc.

Scholars in most academic fields are increasingly using online tools and environments (social media, blogs, online reference managers, etc.) to engage with scholarly literature and other events such as lectures, conferences and symposia (Nicholas and Rowlands 2011). Digital Humanities scholars are not the exception (Ross, Terras, Warwick, Welsh, 2011; Terras 2012), but there is a paucity of bibliometric research regarding the type of publications and impact of those publications that they choose to publish their research. In spite of the extensive work by the Statistical Cybermetrics Group (University of Wolverhampton), digital humanities as a specific field of academic publishing remains largely unexplored. The poster we propose seeks to make a contribution by employing alt-metrics, the quantitative and qualitative "study and use of scholarly impact measures based on activity in online tools and environments" (Adie and Roe, 2013; Cameron 2009; Cronin 2001; Priem et al 2012) to assess the publishing landscape in digital humanities. The data about research online engagement we can obtain from them is discipline-agnostic; it is the online behaviour of researchers and interactions with the outputs from different disciplines what can significantly differ.

The poster seeks to make a contribution to the debate about the role of open access and alternative metrics in contemporary research. The poster will be accompanied by an open access online resource including further analysis and the source data, encouraging fellow researchers to explore, reuse and visualise in different ways. This companion site will discuss how alt-metrics data could potentially contribute to –or eventually generate a culture towards– strengthening the evidence informing impact case studies for journals publishing digital humanities scholarship.

Many questions arise from looking at the data. How can we better understand and use it? Can we classify articles and journals by the kind of attention they get? Are there common patterns between themes? How do they compare to articles and journals in other disciplines? Can online attention metrics encourage specific types of online behaviour amongst digital humanities scholars and across disciplinary? What does it mean if somebody tweets a paper -what's the tweeter trying to do? How can the studied journals maximise the online engagement with the research they publish?

This poster and its companion online site will aim to provide some answers in order to provide recommendations and best practices that might help democratise and increase the international access to peer-reviewed digital humanities research.

References

- Adie, E., and William R.** (2013) "Altmetric: Enriching Scholarly Content with Article-Level Discussion and Metrics." Learned Publishing 26, no. 1 11-17. WEB. http://figshare.com/articles/Enriching_scholarly_content_with_article_level_discussion_and_metrics/105851. Accessed 28 October 2013.
- Cameron N, and Shirley W.** (2009) "Article-Level Metrics and the Evolution of Scientific Impact." PLOS Biology 7, no. 11 e1000242. WEB. <http://www.plosbiology.org/article/info%3Adoi%2F10.1371%2Fjournal.pbio.1000242>. Accessed 28 October 2013.
- Chamberlain, Sc.** (2013) "Consuming Article-Level Metrics: Observations and Lessons." Information Standards Quarterly 25, no. 2 , 4-13. <http://www.niso.org/publications/isq/2013/v25no2/chamberlain>
- Cronin, B** (2001). "Bibliometrics and beyond: Some Thoughts on Web-Based Citation Analysis." Journal of Information Science 27, no. 1, 1-7.
- Nicholas D, Rowlands I.** (2011). Social Media use in the Research Workflow, Information Services and Use 31(1-2, 2011): 61-83
- Ross, M., Terras, C. Warwick, A. Welsh,** (2011) "Enabled backchannel: conference Twitter use by digital humanists", Journal of Documentation, Vol. 67 Iss: 2, pp.214 – 237
- Priem, J., Groth, P. and Taraborelli, D.** (2012). "The Altmetrics Collection." PLOS ONE 7, no. 11 (2012): e48753. WEB. <http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0048753>. Accessed 28 October 2013.
- Statistical Cybermetrics Group.** University of Wolverhampton, United Kingdom. Publications. WEB. <http://cybermetrics.wlv.ac.uk/publications.htm>. Accessed 28 October 2013.
- Suber, P.** (2012). Open Access. (Cambridge, MA: MIT Press). WEB and print. <http://mitpress.mit.edu/books/open-access> Accessed 28 October 2013.
- Terras, M.** (2012). "Infographic: Quantifying Digital Humanities". Melissa Terras' Blog. 20 January 2012. WEB. <http://melissaterras.blogspot.co.uk/2012/01/infographic-quantifying-digital.html>. Accessed 28 October 2013.
- Terras, M.** (2012). "Is blogging and tweeting about research papers worth it? The Verdict." Melisa Terras' Blog. 3 April 2012. WEB. <http://melissaterras.blogspot.co.uk/2012/04/is-blogging-and-tweeting-about-research.html>. Accessed 28 October 2013.

Big Data and the Literary Archive: Topic Modeling the Watson-McLuhan Correspondence

Quamen, Harvey

hquamen@ualberta.ca
University of Alberta

Hjartarson, Paul

phjartar@ualberta.ca
University of Alberta

Introduction

The world of Big Data has introduced humanist scholars to new and relatively unfamiliar data-handling techniques such as data mining, graph visualizations, document clustering, and topic modeling. Topic modeling techniques "are statistical methods that analyze the words of the original texts to discover the themes that run through them, how those themes are connected to each other, and how they change over time."¹

This paper examines how one digital humanities project—the digitization of a literary archive—is using topic modeling in order to help users browse and discover the contents of the archive.

Background and Methodology

Topic models have been commonly used to classify documents^{2 3} or to cluster tagged artifacts⁴; however,

researchers are increasingly using topic modeling as a way to allow users to browse and search large corpora, often through data visualizations.^{5 6 7} Traditional search techniques often fail with large corpora because, as one researcher puts it, "users may not be familiar with the vocabulary that defines the topics of their interest, or simply they may wish to get a broad summary of the collection in order to guide their searches."⁸ If data visualization in the sciences typically happens after the research and number crunching is finished, data visualization in the humanities often serves as a preliminary tool of exploration and discovery. Literary archives are a perfect litmus test: online finding aids often lack detail and there's a storied tradition of having to pay one's scholarly dues by enduring long, physically exhausting archival sessions whose consequences range from bad posture and poor eyesight to contagion and even meningitis.⁹ Digitization and new search techniques can help make archival research more accessible for all scholars.

Our research team is experimenting with a wide variety of big data techniques as we continue a project funded by the Social Sciences and Humanities Research Council of Canada to digitize the archives of Wilfred and Sheila Watson, two 20th-century Canadian writers. The Watson Archive—a lifetime of journals, manuscript drafts, notes, sketches, artwork, newspaper clippings, reviews, and correspondence distributed across two universities 1500 km apart—clearly exceeds the scope of any one publication or scholarly project. In archival terms, Wilfred's papers occupy 10.6 metres of shelf space (producing over 101,000 digital images), while Sheila's occupy a further 8.4 metres, none of which has yet been digitized. Topic modeling is one promising way for scholars to understand the content and the contours of these two separate but related Watson archives in ways that move significantly beyond online finding aids or the serendipity of sifting through the materials in person.

Big data techniques like topic modeling have found a mixed reception in the humanities, however. Twenty-five years ago, Carlo Ginzburg argued that the humanities were different from the sciences because while the humanities privileged "the study of individual cases, situations, and documents precisely because they are individual," the sciences investigated only phenomena that were quantifiable and repeatable.¹⁰ Big Data has begun to challenge that neat division: perhaps most famously, Franco Moretti has used quantitative techniques in order to "distant read" literary history,¹¹ while, more recently, Johanna Drucker has indicted those same methods as "pernicious" because they "violate[e] the very premises of humanistic inquiry."¹²

Research

This paper engages those debates in light of one particular test case, a corpus of 413 letters that Wilfred Watson, Sheila Watson, and Marshall McLuhan wrote to each other over a period of more than twenty years. Because Sheila Watson studied for her PhD under McLuhan's supervision and Wilfred Watson collaborated with McLuhan on the 1970 monograph *From Cliché to Archetype*,¹³ the archival letters range across a wide spectrum of topics from the personal to the professional, from the microdata of timelines, draft revisions, and gossip to the macrodata of history, culture and civilization. Topic modeling offers scholars one browsable entrance into such an unwieldy corpus, a corpus that is nonetheless just a tiny fraction of the entire archive. Our team has built a prototype interface that allows scholars to choose the number of topics into which the algorithm should cluster the letters and then, in the resulting force-directed graph, users are able to click on nodes that reveal the significant words that form each cluster and to browse each cluster's individual letters. Data visualization merges with user interface design to provide scholars a new means of engaging the Watson archive.

Our interface prototype provides scholars a hybrid between a "distant reading" of the archive and a full text search. Broad, long-term patterns and topic shifts become immediately visible. For example, clustering the letters into as few as three or four groups reveals that, as McLuhan and co-author Wilfred Watson

collaborated on *From Cliché to Archetype*, their paradigmatic literary figure shifted from Wyndham Lewis to James Joyce. The conversations about why Joyce proved more satisfactory than Lewis, of course, appear only in the letters and not in the finished monograph. Through data visualization and topic modeling, however, scholars can explore how ideas shift and change over time and can "zoom in" to important moments in the corpus of letters.

Our argument, then, is that topic modeling and other big data techniques are increasingly invaluable to humanists, especially as scholars are confronted by the overwhelming data available in even the most modest-sized archives. Topic modeling provides humanist scholars a valuable new way to explore large collections of texts and artifacts—not necessarily to determine definitively or algorithmically how texts should be classified or clustered, but as experimental, dynamic means of seeing patterns of similarity and difference in a wide range of materials. The result is that humanists are now increasingly able to move beyond the individual, idiosyncratic cases that Ginzburg described to see how people and ideas and discourses change over time.

References

1. Blei, David M. (2011) *Introduction to Probabilistic Topic Models*. www.cs.princeton.edu/~blei/papers/Blei2011.pdf . p.2.
2. Zhou, Shabin, Kan Li and Yushu Liu (2009). *Text Categorization Based on Topic Model*. International Journal of Computational Intelligence Systems 2.4 (December 2009): 398-409.
3. Song, Min, and Su Yeon Kim (2013). *Detecting the Knowledge Structure of Bioinformatics by Mining Full-Text Collections*. Scientometrics 96: 182-201.
4. García-Plaza, Alberto Pérez, Arkaitz Zubiaga, Víctor Fresno, and Raquel Martínez (2012). *Reorganizing Clouds: A Study on Tag Clustering and Evaluation*. Expert Systems with Applications 39: 9483–9493.
5. Shao, Jian, Shuai Ma, Weiming Lu, and Yueling Zhuang (2012). *A Unified Framework for Web Video Topic Discovery and Visualization*. Pattern Recognition Letters 33: 410–419.
6. Anaya-Sánchez, Henry, Aurora Pons-Porrata, and Rafael Berlanga-Llavori (2010). *A Document Clustering Algorithm for Discovering and Describing Topics*. Pattern Recognition Letters 31: 502–510.
7. Gretarsson, Brynjar, John O'Donovan, Svetlin Bostandjiev, Tobias Hollerer, Arthur Asuncion, David Newman, and Padhraic Smyth (2012). *TopicNets: Visual Analysis of Large Text Corpora with Topic Modeling*. ACM Transactions on Intelligent Systems and Technology 3.2 (February 2012): Article 23. 26pp.
8. Anaya-Sánchez (2010), p. 502.
9. O'Driscoll, Michael and Edward Bishop (2004). *Archiving 'Archiving.'* English Studies in Canada 30.1 (March 2004): 1-16.
10. Ginzburg, Carlo (1989). *Clues: Roots of an Evidential Paradigm*. In *Clues, Myths, and the Historical Method*. Trans. John and Anne C. Tedeschi. Baltimore: Johns Hopkins, UP. 96-125.
11. Moretti, Franco (2005). *Graphs, Maps, Trees: Abstract Models for Literary History*. London: Verso.
12. Drucker, Johanna (2011). *Humanities Approaches to Graphical Display*. Digital Humanities Quarterly 5.1. www.digitalhumanities.org/dhq/vol/5/1/000091/000091.html
13. McLuhan, Marshall and Wilfred Watson (1970). *From Cliché to Archetype*. NY: Viking.

The Digital Alchemist: A Mixed Reality Exploration of Jonson's Alchemist as Site-Specific Theatre

Quinsland, Kirk

Fordham University, United States of America

Rouse, Rebecca

Rensselaer Polytechnic Institute, United States of America

Introduction

The Digital Alchemist (DA) is an interdisciplinary project currently in development that is dedicated to presenting both scholars and a general audience with tools for understanding Ben Jonson's *The Alchemist* as a piece of site-specific theatre (digitalalchemist.weebly.com/). Our aim is for the DA to provide a framework for the development of multi-dimensional digital environments that explore works of literature in the historical contexts of their original production.

The Alchemist and Site-Specificity

Jonson's *The Alchemist* is an ideal piece for exploring the relationships between place and performance. The narrative of the play is set entirely within London's Blackfriars district, where Jonson himself lived, and where the play was first performed in the Blackfriars Theatre. Using a variety of digital technologies, we are able to recreate aspects of what it would have been like to see the play in its original context, as well as articulate the argument for understanding *The Alchemist* as a piece of site-specific theatre.

The term site-specific is most often used in relationship to visual arts such as sculpture, although more recently performance studies has claimed the term as well [Kaye (2000); Pearson and Shanks (2001); Wilkie (2012)]. Work in other fields on the concepts of space and place has also provides relevant connections to the topic of site-specificity and locative artifacts: from cultural studies [de Certeau, M. (1984); Bachelard, G. (1994); Auge, M. (1995), Kwon, M. (2002)], literary studies [Bly, M. (2007); Hopkins, D.J. (2007); Howard, J. (2009)], architecture [Tuan, Y. (1977)], design [Charitos, D. (2009)], and media studies [MacIntyre, B. et. al. (2004); Bolter, J. D. et. al (2006)]. Across all of these engagements with the concept of site, a deep interconnection between place and content is emphasized. Sculptor Richard Serra has provided what continues to be a key definition for site-specific art: "Site-specific works deal with the environmental components of given places. The scale, size, and location of site-specific works are determined by the topography of the site, whether it be urban or landscape or architectural enclosure" [Serra in Kwon (2002), 12]. A site-specific work, then, is one that is created specifically for a particular location, and whose content is determined by the space it will occupy.

A close examination of Jonson's *The Alchemist* and supporting historical documentation suggests this work includes a deep connection between place and content, and was likely intended to be understood by its contemporary audiences as site-specific. The play was written for the Blackfriars Theatre, and the action of the play was set in the Blackfriars district as well, in a house near to the theatre itself. This positioning is unique in the early modern theatrical world: while many plays were written to be performed in the Blackfriars Theatre, *The Alchemist* is one of only a small handful that is also set in that environment. Literature scholar R. L. Smallwood explains that Jonson's "[...] temporal and locative immediacy is carefully built into the play. The audience within the play and the audience in the theater are one and the same" [157]. Despite these and other strong textual ties to site, the fact that the site of the play's original setting and performance no longer exists creates a disruption in the relationship between the play and the site. The DA seeks to address this disconnect.

The Problem of Representation: Site, Authenticity and Simulation

Even though this project is developed around the concept of site-specificity, there is a problem in determining exactly

what constitutes a "site." This problem of the shifting nature of the term has been explored by Kaye in depth. However, this problem becomes magnified in the case of *The Alchemist* in the context of creating a map of the Blackfriars site based on the pertinent historical documentation contained in the Loseley manuscript archive. This archive includes documents spanning nearly two hundred years (1489-1682), during which the Blackfriars changed in many significant ways. Nevertheless, some important aspects of the site persist even today: most of the district's original streets are in their original locations [Whitfield, P. (2007); Barber, P. (2012)], and a few of them--most notably Playhouse Yard--reference topographical features that have long since vanished.

Because of the problematic nature of determining "the site," the DA must contend with a kind of temporal-specificity that parallels site-specificity: if *The Alchemist* was written for a particular place in London, so too was it written for a particular time. We can never completely replicate an environment, or fully recapture the audience experience of the original production. It is possible, however, with the use of digital media, to simulate the environment in a way that engages contemporary audiences with the site-specific nature of the play.

Additionally, the use of digital technology itself results in complex negotiations around the concepts of aura and authenticity. This question of aura becomes especially interesting in the case of work that combines the physical and digital, as in augmented or mixed reality [Bolter, J. D. et. al (2006)]. As opposed to virtual reality, mixed and augmented reality systems maintain a connection to the physical. This connection can provide the possibility for a differently embodied form of experience, which includes elements of a reproductive technology experience, and is "both immediate and mediated" [Bolter 29], thus highlighting both the presence of site (or a site's remains) as well as the absence of site.

Solution(s): Multi-nodal Experience Design

The DA addresses the problematic nature of site-specificity in a site that no longer exists by providing multiple experiences, addressing different aspects of the play and the remaining traces of the site. These multiple nodes provide different kinds of access for multiple audiences. Each of these nodes provides a means for audiences to engage and interact with the archive itself in different ways depending on participant interest. For example, scholars will be able to explore the original documents along with their connections to Jonson's text, while participants in an augmented reality experience will be able to interact with a version of the play unlike any previously staged. These different experiences are not mutually exclusive, and in fact compelling connections are revealed by participating in more than one, although this is not required.

Conclusions

Moving forward, we plan to conduct user testing to solicit feedback on the each node of the project, and develop a framework that could be applied in practice to create similar explorations of other literary works that also demand attention to site-specificity. Future objectives include collaborations with colleagues in a variety of fields, such as architectural history and palaeography, to expand our design such that the content is accessible for a range of interdisciplinary uses.

References

- Auge, M.** (1995). *Non-Places: Introduction to an Anthropology of Supermodernity*. London and New York: Verso.
- Bachelard, G.** (1994). *The Poetics of Space*. Boston: Beacon Press.
- Barber, P.** (2012). *London: A History in Maps*. London: The British Library.

- Benjamin, W.** (1968). "The Work of Art in the Age of Mechanical Reproduction." In Walter Benjamin: Illuminations. Trans. H. Zohn. pp. 155-200. New York: Schocken Books.
- Bly, M.** (2007). "Playing the Tourist in Early Modern London." PMLA, 122:1, pp. 61-71.
- Bolter, J. D., MacIntyre, B., Gandy, M., Schweitzer, P.** (2006). "New Media and the Permanent Crisis of Aura." Convergence, 12:1, pp. 21-39. London: Sage Publications.
- Charitos, D.** (2009). "Precedents for the Design of Locative Media." In Future Interaction Design II. Eds. P. Saariluoma and H. Isomaki. pp. 141-156. London: Springer.
- de Certeau, M.** (1984). *The Practice of Everyday Life*. Trans. S. Rendall. Berkeley: University of California Press.
- Hopkins, D.J.** (2007). *City/Stages/Globe: Performance and Space in Shakespeare's London*. London and New York: Routledge.
- Howard, J.** (2009). *Theater of a City: The Places of London Comedy, 1598-1642*. Philadelphia: University of Pennsylvania Press.
- Kaye, N.** (2000). *Site-Specific Art: Performance, Place and Documentation*. London and New York: Routledge.
- Kwon, M.** (2002). *One Place After Another: Site-Specific Art and Locational Identity*. Cambridge MA: MIT Press.
- MacIntyre, B., Bolter J. D., Gandy, M.** (2004). "Presence and the Aura of Meaningful Places." Presence, October 2004, pp.13-15.
- Pearson, M., Shanks, M.** (2001). *Theatre/Archaeology*. London and New York: Routledge.
- Smallwood, R. L.** (1981). "Here, in the Friars': Immediacy and Theatricality in *The Alchemist*." Review of English Studies. 32:126.
- Tuan, Y.** (1977). *Space and Place: The Perspective of Experience*. Minneapolis: University of Minnesota Press.
- Whitfield, P.** (2007). *London: A Life in Maps*. London: The British Library.
- Wilkie, F.** (2012). "Site-specific Performance and the Mobility Turn." Contemporary Theatre Review, 22:2, pp. 203-212. London: Routledge.

How to make games more GLAMorous: developing game prototypes for the museum and cultural heritage sector in India

Ray Murray, Padmini
 padmini.raymurray@stir.ac.uk
 University of Stirling

1. Introduction

1.1. Overview

Meghdoot: Using new technologies to tell age-old stories was conceived as part of the Arts and Humanities Council sponsored 'Unplay' project, under the aegis of its Unbox scheme, a collaborative endeavour with the British Council and the Unbox Festival, an interdisciplinary festival that celebrates creative work at the intersections of art, design and technology.

1.2. Methodology

Despite the exemplary progress in areas of computing, programming and animation in India, there are still no major game companies or developers creating games that compete on a global level¹. In order to explore why this might be, and how our research might stimulate such growth, we are using a two-pronged approach; our methodology can be split into data gathering and analysis of that data, and practice-based

research through the creation of a game. The data-gathering is being conducted by an online and offline survey of game developers and gamers, across socio-cultural and linguistic communities in India (covering platform reach, importance of storytelling, gaming practices, problems of accessibility) and through interviews with game developers, gamers, critics, game companies and studios in the UK, US and India. Scoping the market needs to be underpinned by empirical research, which is one of the aims of this project, these explorations will look both inwards at the Indian market, but also the potential of videogames made in India in the global market.

Meghdoot was formulated by me in my role as principal investigator as a response to the open-ended brief articulated by my research partners, which was to create a Kinect-based game while based in India, working with local partners. My conceptualisation of the game was underpinned by three assumptions: that the affordances of the Kinect should be used to its fullest potential; that we should make a conscious effort to move away from Anglo-Saxon linear narrative sequences in the game's design; and that while using deploying an aesthetic that was inspired by the game's Indian origins, it should not resort to usual tropes of the exotic or the oriental. The narrative of the game itself operates on a meta-textual level, in that it is a videogame (a storytelling medium) about storytelling. In order to make the best use of the affordances of the Kinect, the three levels of the game draws on narrative modes that co-exist simultaneously in contemporary India: the textual, the gestural and the oral. As the first phase that developed *Meghdoot* was successful, my research partners and I have received a second tranche of funding which has allowed us to develop another game, drawing on our learnings obtained from our earlier collaboration.

Research for both games involve modelling artefacts within the game on heritage artefacts from museums, thus providing alternative channels of interpretation for the cultural heritage sector in India, which till date has been quite conservative and prescriptive. These assets have been developed by 3D scanning and rendering cultural objects from various institutions, and part of our research focuses on exploring how the narratives embedded in these objects can be communicated in the gameplay and communication and work as supplementary collateral. The aim is to ultimately educate players in basic programming, in order to create very simple 3D controllers that can be used in the game. In order to do this effectively, empirical research will be conducted into levels of literacy required to acquire simple digital skills—a feat that has been made infinitely more cost-effective and feasible by the introduction of such mini computers as the Raspberry Pi on the market.

The aims of this project are perfectly fitted to this year's conference theme of digital cultural empowerment. By incorporating a 'making' aspect to the game, this research will address how hands-on interaction with code and technologies can enhance the experience of the game as it can create a sense of ownership of artefacts used with the game. The attractiveness of 'digital making' and its relationship to games, in both 2D and 3D space and how it can educate users, has been an important focus of recent research, due to games like Minecraft. Transposing these practices to an Indian context, should yield significant outputs as "Technology usage in turn is shaped by the socioeconomic location of the user, especially in regards to gender and caste"². This poster session will demo both game prototypes, and be enhanced by slides of storyboards and showing stages of research design. Discussion of the research will focus on how affordances, cultural context, market and the target demographic should shape game design.

References

1. E.W. Adams (2009), *The Promise of India: Ancient Culture, Modern Game Design* at the NASSCOM Animation and Gaming Summit 2009, Game Development Summit Keynote Address, November 7, 2009. [<http://www.designersnotebook.com/Lectures/India/India.htm>]
2. A. Schwittay (2011), *New Media Practices* in India: Bridging Past and Future, Markets and Development.¹

International Journal of Communication, Volume 5, (2011) pp. 349-379.

S. Mukherjee (2012), *India in 'Encyclopedia of Video Games: The Culture, Technology, and Art of Gaming, Volume 1'* edited by Mark J.P. Wolf (Santa Barbara: Greenwood, 2012) pp. 312-314.

Enhancing Access to Online Oral History: Oral history in the Digital Age (OHDA) and Oral History Metadata Synchronizer (OHMS)

Rehberger, Dean

deanreh@gmail.com
Michigan State University

Boyd, Douglas

douglasaboyd@gmail.com
University of Kentucky

Oral history is in a profound transition, from an extensive period when sophisticated technology meant utilizing tape cassettes, to a time when the field has moved into the digital, networked, multi-media rich age. The transition into a digital world, and the flexibility it brings, has changed the costs of doing oral history, standards of practice and scholarship, and the vehicles for access. Resulting issues are deeply complex and often dynamic. Digital video is now readily affordable, but the field remains deeply divided over its use and role. Equally important, the digital age makes widespread access and use of both audio and video oral narratives, as well as transcripts, increasingly affordable, but it also highlights major questions about intellectual property rights and informed consent. The Oral History in the Digital Age (<http://ohda.matrix.msu.edu>) attempts to address many issues faced with new technologies.

OHDA has over 72 essays, 12 videos, and many other resources, including an interactive review of recording equipment called "Ask Doug," a large collection of online oral history collections, and consent forms. Oral History in the Digital Age (OHDA) is a product of an Institute of Museum and Library Services (IMLS) National Leadership project and a collaboration among the Michigan State University Museum; Michigan State University Digital Humanities Center, Matrix; the American Folklife Center (AFC/LOC), the Library of Congress; the Smithsonian Center for Folklife and Cultural Heritage (CFCH); the American Folklore Society (AFS); the Louie B. Nunn Center for Oral History, University of Kentucky Libraries; and the Oral History Association.

In development with OHDA is the Oral History Metadata Synchronizer (OHMS). The Louie B. Nunn Center for Oral History at the University of Kentucky Libraries has created a web-based, system called OHMS (Oral History Metadata Synchronizer) to inexpensively and efficiently enhance access to oral history online. OHMS provides users word-level search capability and a time-correlated transcript or index connecting the textual search term to the corresponding moment in the recorded interview online.

In 2011, the Institute of Museum and Library Services (IMLS) awarded the Nunn Center a \$195,853 National Leadership Grant to further develop their Oral History Metadata Synchronizer (OHMS). The grant project is designed to prepare OHMS for open source distribution and to create compatibility between OHMS and other popular content management systems empowering institutions, both large and small, to provide an effective, user-centered discovery interface for oral history on a large scale. In addition to developing OHMS compatibility with open source content management systems such as OMEKA and KORA, and larger scale commercial systems such as CONTENTdm, this project has developed multimedia tutorials instructing users on the use, installation and deployment of OHMS within particular content management systems.

Beginning in 2013, the open source OHMS system has integrated an interview-indexing module. Instead of relying solely on transcription, OHMS can now be utilized to create segment-level metadata that correlates to the corresponding moment in the recorded online interview. The new interview-indexing module opens up new capabilities for OHMS and expands possibilities for quickly providing enhanced access to far more interviews online for a fraction of the price of transcription.

Marked E-Books and Kindle's popular highlight culture

Rowberry, Simon

University of Winchester, United Kingdom

The current project analyses the evidence of readership available through the public facing popular highlights feature of Amazon's Kindle platform. In order to be considered a popular highlight, the text must be shared by three users. There are over one million quotations that meet this basic criteria and can be analyzed in similar ways to evidence of marginalia and provenance in book historical research. The present research analyses the popular highlights as a measure of various genre's popularity as well as observing usage patterns of the highlighting and sharing features.

The static e-book has become embedded in the public's imagination as an exemplar of the future of reading on the screen. The Kindle is one of the forerunners in the commercial e-book marketplace, encompassing a range of both software and hardware platforms and offering millions of titles. While others have begun to explore the impact of e-book culture, (Galey 2012; Lang 2012; Wu 2013; Thomas & Round 2013), the current project focuses on the traces readers leave directly on their Kindles. Amazon offer tools to share annotations and highlights of their eBooks to replicate print marginalia. The data for popular highlights is shared on a public-facing webpage (Amazon.com, Inc. 2013) that can be collected for analysis. This research offers an approach to the empirical study of reception on a previously unprecedented scale and offers an insight into what users find interesting about the material they are reading.

The data was collected using wget on the Kindle Popular Highlights website, as Amazon.com does not currently offer an API for the dataset. The project focused on the Popular Highlights feature and the metadata pertaining to the book title, author, quotation and number of highlights. While this does not provide evidence of individual readers, it can be used to analyse patterns of readership and marginalia. An initial foray produced the first 100,000 popular highlights (out of a dataset of over 1,000,000 highlights) that were produced by over 8 million shared highlights. Unfortunately this method left many artefacts when converting certain characters, so the data was cleaned and organized.

The initial results revealed some interesting patterns. The most highlighted books were primarily Young Adult (YA) fiction, literary classics, pop science and self-help. Individual passages can be highlighted more than 1,000 times, with a quotation from Catching Fire (The Second Book of the Hunger Games) received over 17,000 highlights. Each genre's annotations often fit into roughly categorized groups: literary classics and pop science produce pithy aphorisms; self-help books are quoted for their instructions; and YA generally highlighted "spoilers" and dialogue that is central to the novel's plot. Over 90% of the quotations are under 350 characters, although occasionally readers will highlight a whole page. Since one of the core features of the popular highlight function is the ability to re-use the quotations as tweets, brevity of quotation length is expected and confirmed as 42% of the highlights are tweetable. As the number of highlights fall, the books' genres tend to become more esoteric and the highlights become fuzzier. Some of these bear the marks of experimenting with the feature or more

playful purposes, such as "THE" in the New Oxford American Dictionary receiving 73 highlights.

The analysis comes with a few caveats: (1) the Kindle is only one eBook provider and is not representative of digital reading; (2) it is unknown to what degree this data is representative of reading on the Kindle in general; (3) the data does not currently include 90% of the data; and (4) without a finer breakdown of the users' demographics, the data can only tell us so much about what the readers are attempting to do through highlighting. Nonetheless, the Kindle Popular Highlights dataset offers a snapshot into the possible ways in which book historical research can be conducted in the early twenty-first century.

References

- Amazon.com, Inc.**, 2013. *Most Highlighted Passages of All Time*. Available at: https://kindle.amazon.com/most_popular/highlights_all_time/.
- Gale, A.**, 2012. *The Enkindling Reciter: E-Books in the Bibliographical Imagination*. Book History, 15(1), pp.210–247.
- Lang, A. ed.**, 2012. *From Codex to Hypertext: Reading at the Turn of the Twenty-First Century*, Amherst and Boston: University of Massachusetts Press.
- Thomas, B. & Round, J.**, 2013. *Digital Reading Network*. Available at: <http://www.digitalreadingnetwork.com/> [Accessed October 27, 2013].
- Wu, Y.-H.**, 2013. *Kindling, Disappearing, Reading*. , 7(1). Available at: <http://www.digitalhumanities.org/dhq/vol/7/1/000115/000115.html> [Accessed October 27, 2013].

Sustainability?! Four Paradigms for Humanities Data Centers

Sahle, Patrick

sahle@uni-koeln.de

Cologne Center for eHumanities, University of Cologne

Kronenwett, Simone

simone.kronenwett@uni-koeln.de

Data Center for the Humanities, University of Cologne

Blumtritt, Jonathan

jonathan.blumtritt@uni-koeln.de

Data Center for the Humanities, University of Cologne

Introduction

The Digital Humanities are about creating knowledge. Methods and tools, formats and data structures are designed to generate valuable digital research data and accessible resources. However, these products of scholarship are created within the framework of temporary projects. So what happens when projects come to an end? What about sustainability and future accessibility of results which were produced on cost of public money? One common and obvious answer of our times concerning the questions raised above might be 'long-term preservation'. The archiving of digital data has been under research for many years now and will lead to the establishment of generic methodological and technical solutions as well as institutionalized data archives. But is this really the full answer to the questions posed? Typically, the goal of research projects in the humanities is not only generating 'records', 'data objects' or 'collections' that follow standardized models and formats and are thus easily stored together and reused by others. Rather, research in the humanities is characterized by individual, local approaches and data models that lead to very specific databases and, more importantly, publications with their own logic of presentation, access and usability.

Data, Resources, and Data Centers

Within the humanities, there is an important difference between 'data' and 'resources'. This distinction has to be taken into account: often data is usable only within its context and is made accessible by specific web presentations. But who will take care of these 'resources' in the long run? On the one hand, long-term maintenance for sometimes idiosyncratic 'living systems' cannot be ensured by individual scholars - not even by academic research departments. On the other hand, nonspecific libraries and archives - including data archives - cannot be expected to provide knowledge on data models and standards that are specific within the humanities. For this data, as well as for the particular 'resources', a dedicated research data management is needed as well as workflows and business models for the perpetuation of the presentational systems. A comprehensive solution to these problems must be built upon institutions that can make a permanent commitment. These institutions could be called 'data centers for the humanities' and can be attached to other institutions like libraries, archives, computing centers, academic departments or faculties - or digital humanities centers.

Four Paradigms for Humanities Data Centers

These data centers have to provide special research data management for humanities research projects - ideally right from the beginning - to ensure archivability, accessibility, reusability, maintainability and visibility for a long time. The wide range of their tasks and goals can be described in analogy to four paradigms which we already know from our traditional research and information ecosystem:

1.) Archive paradigm. Research data has to be archived in order to be permanently preserved and secured. As pure 'data', as 'bits and bytes', research products in the humanities are not very specific and may thus be stored in generic data archives on which a data center may rely as part of the more basic infrastructure. Here, the integrity of the data has to be ensured, data may be converted into different formats, and data objects may be delivered for reuse when needed.

2.) Library paradigm. A library maintains a descriptive catalog of its holdings, cares for unique call numbers and keeps data ready for direct access. In terms of digital infrastructures this means caring for metadata, persistent identifiers and technical interfaces (such as APIs), and the integration of data and metadata into dedicated portals that allow browsing and searching.

3.) Museum paradigm. The common use case of digital output in the humanities is not to harvest or to integrate third party data via APIs. Rather, it is to consult the digital publication that is approachable and readable for humans. Digital research objects have to be presented. Often very special websites and portals are created within research projects and have to be kept alive. As in museums, important holdings are presented in a permanent exhibition. But the exhibition may also be changed, reconfigured and redone over the course of time as well.

4.) Workshop paradigm. Digital libraries evolve to virtual research environments. More and more, the presentation of digital objects and active work on the data coincide. When research platforms pass over from a project to a data center, the work should not have to stop. In an ideal world a data center will also maintain current research environments and keep them alive and working. With either generic or dedicated tools and interfaces, a data center will provide important components for the ongoing editing, enrichment and processing of digital research data.

From Theory to Practice

The issues raised here have to be addressed on a methodological and a theoretical level. But the problem of short term projects that become orphaned and whose data is at risk is acute. Therefore, the University of Cologne founded a Data Center for the Humanities in 2012. This paper will report on the theoretical background, further concepts and plans as well as the first practical steps that have already been taken.

References

- Ball, A.** (2012). *Review of Data Management Lifecycle Models* (version 1.0). REDm-MEDProject Document redm1rep120110ab10. Bath/UK: University of Bath. opus.bath.ac.uk/28587 .
- Blanke, T., M. Hedges** (2013). *Scholarly Primitives. Building Institutional Infrastructure for Humanities e-Science*. In: Future Generation Computer Systems. 29/2. 654-661. linkinghub.elsevier.com/retrieve/pii/S0167739X11001178 .
- Büttner, S., H.-C. Hobohm, and L. Müller** (eds) (2011). *Handbuch Forschungsdatenmanagement*. Bad Honnef: Bock +Verchen.
- Burrows, T.** (2011). *Sharing Humanities Data for e-Research. Conceptual and Technical Issues*. In: Sustainable data from digital research. Humanities perspectives on digital scholarship. Proceedings of the conference held at the University of Melbourne. 12-14th December 2011. Sydney: Custom Book Centre. ses.library.usyd.edu.au/handle/2123/7938 .
- Data Center for the Humanities (DCH)**. www.dch.uni-koeln.de .
- Hügi, J., R. Schneider** (2013). *Digitale Forschungsinfrastrukturen in den Geistes- und Geschichtswissenschaften*. Genf: Haute école de gestion de Genève.
- Molloy, L.** (2011). *Oh, the Humanities! A Discussion About Research Data Management for the Arts and Humanities Disciplines*. In: JISC MRD - Evidence Gathering 2011. mrdevidence.jiscinvolve.org/wp/2011/12/16/oh-the-humanities-a-discussion-about-research-data-management-for-the-arts-and-humanities-disciplines/ .
- Neuroth, H., S. Strathmann, and A. Oßwald et al** (eds) (2012). *Langzeitarchivierung von Forschungsdaten. Eine Bestandsaufnahme*. Boizenburg: Werner Hülsbusch.
- Sahle, P., S. Kronenwett** (2013). *Jenseits der Daten: Überlegungen zu Datenzentren für die Geisteswissenschaften am Beispiel des Kölner 'Data Center for the Humanities'* In: LIBREAS. Library Ideas. 23. http://libreas.eu/ausgabe23/09sahle/ .
- Van den Eynden, V., L. Corti, and M. Woppard et al** (2011). *Managing and Sharing Data 2011: Best Practice for Researchers*. Colchester: UK Data Archive. www.data-archive.ac.uk/media/2894/managingsharing.pdf .
- M. Thaller et al** (2011). *DA-NRW: a distributed architecture for long-term preservation*. Proceedings of the 1st International Workshop on Semantic Digital Archives. www.danrw.de/wp-content/uploads/paper13.pdf .
- UK Data Archive**. www.data-archive.ac.uk .
- See discussion on Digital Medievalist Mailing List. Subject *How to make your data live forever (and maybe your project)*. 21-06-2013. www.digitalmedievalist.org/index.html .
- There are different data management lifecycle models. An overview is given in A. Ball (2012). Review of Data Management Lifecycle Models (version 1.0). REDm-MEDProject Document redm1rep120110ab10. Bath/UK: University of Bath. opus.bath.ac.uk/28587/ .

Euterpe's Hidden Song: Patterns in Elegy

Scheirer, Walter

wscheirer@fas.harvard.edu

Center for Brain Science, Harvard University

Forstall, Christopher

forstall@buffalo.edu

Department of Classics, State University of New York at Buffalo

1. Introduction

It is widely acknowledged that classical Latin poetry was heavily influenced by Greek oral poetic traditions. The perception of sound is critical to the experience of the poetic work, and lyrical virtuosity is its hallmark. As Ezra Pound reminds us, "poetry begins to atrophy when it gets too far from music"¹. Recent scholarship in Latin poetry emphasizes the primacy of meter as a framework to organize sound². Classical notions of sound differ, however, with Plato and Aristotle suggesting that music itself is far more organic – a reflection of natural sounds and emotion that emanates deep within the self^{3 4}. In our own study of the role of sound in poetry, we have found patterns in the elegiac form at extraordinarily fine levels, only recently detectable with the use of methods from the digital humanities, and never remarked upon by any ancient or contemporary commentator. These patterns are a signature of the process that generates language, forms meter, and enables the creative act.

2. Aspects of Poetic Composition

Consider the phoneme, which is the fundamental building block of language. Philodemus of Gadara posits that the formation of poetry is contingent upon an agreeable arrangement of phonetic units⁵. Views on how such an arrangement could come to be varied in classical times. *Anomalists* like Varro argue for an injection of novelty into language to draw our attention, while *Analogists* including Julius Caesar recommend an adherence to a perceived natural ordering of linguistic elements⁶. In all of these discussions, the relationship between basic elements is key.

Saussure assumes that differential relationships between phonemes are necessarily negative and reduced to opposition⁷. Deleuze provides a critique of this based on the deeply plausible possibility that a positive relationship can exist. He states, "For opposition teaches us nothing about the nature of that which is thought to be opposed. The selection of phonemes possessing pertinent value in this or that language is inseparable from that of morphemes as elements of grammatical constructions"⁸. An interplay between *reciprocal* sounds leads one to consider alternatives to Saussure's theory.

The linguist Gustave Guillaume writes of a hidden nature of language formulation in the mind before speech actualization⁹. Guillaume argues that potential phonetic, lexical, and semantic forms interact in myriad combinations, defining a "metalinguage" that supports natural language. Deleuze seizes upon this idea, stating that metalinguage "cannot be spoken in the empirical usage of a given language, but must be spoken and can be spoken only in the poetic usage of speech"¹⁰. It is this profound and fundamental insight that we scrutinize in this work.

Is there some way to resolve the opposition between the linguistic theories of Saussure and Guillaume? After all, both systems hint at a probabilistic model that follows some set of rules to enforce cognitive economy, which facilitates memory. Successful "linguistic throws of the dice"¹¹ yield novel turns of phrase, while other combinations fall flat. The link from theory to experimentation that we pursue can be stated as follows: language is constructed by complex unseen interactions between linguistic elements in the mind (Saussure and Guillaume); poetry is unique, in that it provides a window into the aforementioned process (Deleuze). Based on our analysis of a large corpus using the statistical techniques of distant reading^{12 13}, this model appears to have some explanatory power.

3. The Elegiac Form

As a first case study, we examine the poetic form of the elegiac couplet under the lens of the above model. The elegiac meter is used for a variety of themes, most notably erotic love¹⁴. The elegiac couplet is a pair of two different one-line "verses". The first line is identical to dactylic hexameter; the second, often called the "pentameter" line of the couplet, is shorter by two half-feet. The scansion is shown in Fig. 1. In our analysis,

we consider all of the extant elegies from Catullus, Propertius, Tibullus, Ovid, and Martial. A selection of non-elegiac poems is also considered for comparison. All texts come from the Perseus Digital Library¹⁵.

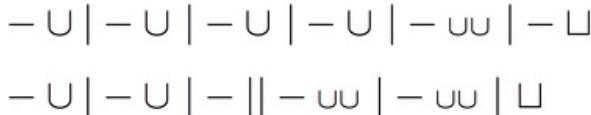


Fig. 1: The elegiac couplet. "—" represents a long syllable, "u" a short syllable, "U" either one long syllable or two shorts, and "U" either one long syllable or a short.

4. Statistical Analysis

For stylistic analysis, the choice of style marker is important^{16 17 18 19}. In this work, we look at a particular form of character-level bi-grams as a proxy for phonemes. Unlike phonemes, character-level bi-grams do not suffer from potential errors introduced by scholarly judgment of how an ancient word might have sounded. A functional bi-gram^{20 21}, when applied at the character-level, is an n-gram-based feature²² that describes the most frequent sound-oriented information in a text. Similar to function words, functional n-grams are those n-grams that are elements of most of the lexicon, necessitating their use. We computed functional bi-grams using a set of custom perl scripts, and calculated statistics via Microsoft Excel. All code and data will be released at DH2014.

In Latin elegies, the most frequent bi-gram is "er". Because of its frequency, this sound alone is quite sensitive to meter, author, and literary era. Other sound choices exert influence upon "er", yielding interesting patterns. It is natural to ask why this is so. In order to see any patterns, we need a large enough feature sampling – otherwise we are simply lost in the noise. This is exactly what a functional bi-gram like "er" is meant to provide.

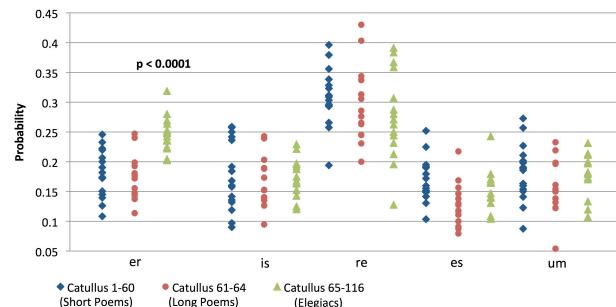


Fig. 2: Functional bi-gram probabilities for Catullus.

Calculating the associated probabilities for "er" over a collection of 50 line samples spanning the entire Catullan corpus exposes a substantial divergence between the polymetric poems, numbered 1-64, and the elegiacs, numbered 65-116, in Fig. 2. The probability of "r" directly following "e" is much higher in the elegiacs. Given just "er" for a sample, it is not difficult to guess its meter. In a rigorous statistical sense, we can ask if the bi-gram "er" is truly significant compared to the other most frequently occurring functional bi-grams. The null hypothesis states that any bi-gram should occur with roughly the same frequency across the entire corpus. For a multiple comparisons scenario with 100 hypothesis tests, the significance level is 0.0001. "er" is statistically significantly different ($p < 0.0001$) via the two-tailed paired t-test.

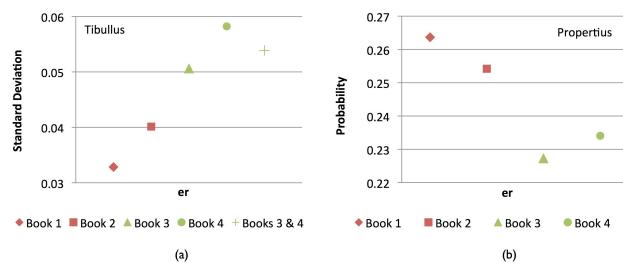


Fig. 3: a. Variation between books 1 & 2 and 3 & 4 of Tibullus; b. "er" bi-gram probabilities for Propertius.

What does "er" tell us about other elegists? In the case of Tibullus, books 3 & 4 are generally acknowledged to be the work of other poets²³, including the perhaps apocryphal contributions of Sulpicia²⁴. Considering variation with respect to "er", there is a noticeable increase in standard deviation for books 3 & 4, which we would expect for multiple authors. In another example, poets after Catullus often end the pentameter line with a two-syllable word²⁵. Propertius, following Catullus and Tibullus, does not do this in books 1 & 2, but adopts the style in books 3 & 4. From the change in values taken on by the "er" feature, a conclusion can be drawn that this new constraint affects sound choice by increasing the number of sound combinations that are used, thus decreasing the probability of "er".

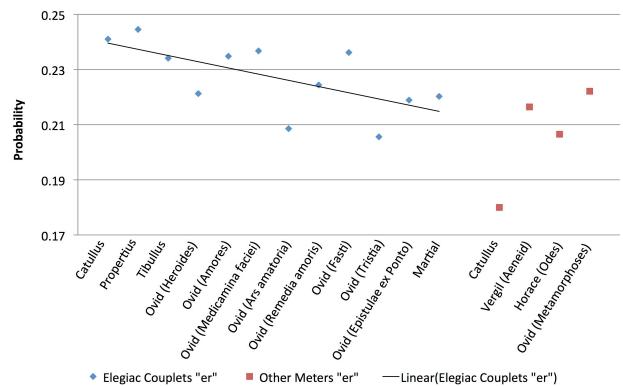


Fig. 4: As the elegiac form evolves, the probability of "er" occurring declines. A linear regression model was fit to the elegists, highlighting the downward trend. The x-axis is arranged chronologically.

A poem's place in time also reflects its style. When viewed chronologically in Fig. 4, the probability of "er" occurring declines. Even within just the Ovidian corpus, this trend is evident. Beyond the change in the pentameter line endings, the pursuit of innovation means more poets look to new sound combinations, which drives down the probability of "er" – and gives some credence to the anomalist argument. Ovid, a master stylist, does not want to sound like his old self as his work progresses. In this vein, a further note can be provided on the magnitude of the meter's role in composition. We find in Fig. 5 a smaller number of longer words in dactylic hexameter compared to elegiac hexameter over the complete works in these meters for thirteen poets. We can explain this as a blending of a genre-dependent signal with the meter signal: the pentameter line tends to have shorter words because of constraints imposed by the meter; this tendency steers word choice towards shorter words in the hexameter line, even though here, as proven by dactylic hexameters, word length is not so constrained by meter.

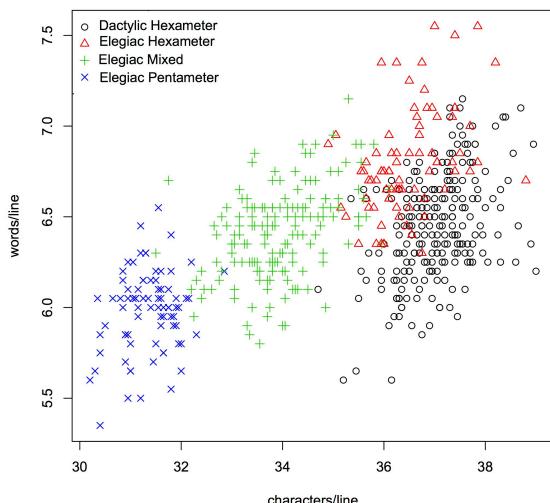


Fig. 5: The interplay between the hexameter and pentameter lines results in shorter hexameter lines with more words, which are atypical outside elegiacs.

5. Discussion

What can be said of these findings? The most frequent sound in a poem provides important clues to the overall construction of language in its aesthetic and historical contexts. Poetry is wonderful for many reasons – that it is a window into the mind is perhaps the most striking.

Acknowledgements

This work was supported by NEH Start-Up Grant Award #HD-51570-12.

References

1. Pound, E. (1960). *ABC of Reading, New Directions*, p. 14.
2. Morgan, L. (2010). *Musa Pedestris: Metre and Meaning in Roman Verse*. Oxford University Press.
3. Grube, G and C. Reeve. (1992). *Plato: Republic*, Second Edition. Hackett Publishing Company, pp. 52-80.
4. Lord, C. (2013). *Aristotle's Politics*: Second Edition. University of Chicago Press, pp. 223-239.
5. D. Armstrong (1995). *The Impossibility of Metathesis: Philodemus and Lucretius on Form and Content in Poetry*. In Obbink, D. ed. *Philodemus and Poetry: Poetic Theory and Practice in Lucretius, Philodemus, and Horace*. Oxford University Press, pp. 210-233.
6. Colson, F. (1919). *The Analogist and Anomalist Controversy*, The Classical Quarterly, 13.1.24-36.
7. Saussure, F. (1966). *Course in General Linguistics*. McGraw-Hill.
8. Deleuze, G. (1968). *Difference and Repetition*. Columbia, p. 205.
9. Guillaume, G. (1984). *Foundations for a Science of Language*. John Benjamins.
10. Deleuze, G. (1968). *Difference and Repetition*. Columbia, p. 193.
11. Ibid., p. 205.
12. Moretti, F. (2005). *Graphs, Maps, Trees: Abstract Models for a Literary History*. Verso.
13. Jockers, M. (2013). *Macroanalysis: Digital Methods and Literary History*. University of Illinois Press.
14. Morgan, L. (2010). *Musa Pedestris: Metre and Meaning in Roman Verse*. Oxford University Press.
15. Perseus Digital Library. Ed. Gregory R. Crane. Tufts University. www.perseus.tufts.edu. Accessed October 17, 2013.
16. Burrows, J.F. (1989). *An Ocean Where each Kind...: Statistical Analysis and Some Major Determinants of Literary Style*, Computers & the Humanities, 23.309-321.
17. Eder, M. (2008). *How Rhythmic is Hexameter: a Statistical Approach to Ancient Poetry*, Digital Humanities, 2007.
18. Juola, P. (2008). *Authorship Attribution*, Foundations and Trends in Information Retrieval 1.3.233-334.
19. Hoover, D. (2013). *The Full-Spectrum Text-Analysis Spreadsheet*, Digital Humanities, 2013.
20. Forstall, C.W. and W.J. Scheirer. (2009). *Features from Frequency: Authorship and Stylistic Analysis Using Repetitive Sound*, Journal of the Chicago Colloquium on Digital Humanities and Computer Science 1.2.1-23.
21. Forstall, C.W., Jacobson, S.L. and W.J. Scheirer. (2011). *Evidence of Intertextuality: Investigating Paul the Deacon's Angustae Vitae*, Literary and Linguistic Computing 26.3.285-296.
22. Jurafsky, D. and J. Martin. (2009). *Speech and Language Processing*: Second Edition. Pearson Prentice Hall.
23. Conte, G.B. (1999). *Latin Literature: A History*. Johns Hopkins, pp. 330-331.
24. Hubbard, T. (2005). *The Invention of Sulpicia*, The Classical Journal, 100.2.177-194.
25. Platnauer, M. (1951). *Latin Elegiac Verse: Study of Metrical Uses of Tibullus, Propertius and Ovid*. Cambridge, p. 17.

DARIAH-DE – Digital Infrastructure for the Arts and Humanities

Schmunk, Stefan

Göttingen State and University Library, Research & Development Department, Germany

Smith, Kathleen

Göttingen State and University Library, Research & Development Department, Germany

Blümm, Mirjam

Göttingen State and University Library, Research & Development Department, Germany

DARIAH-DE supports digitally-enabled research and teaching in the arts and humanities. The project is developing a research infrastructure, which will offer tools, core services, and access to research data as well as materials for research and education in the digital humanities (DH). The central objective of DARIAH-DE is to enable the interoperability of tools and research data. Following internationally valid and accepted standards and policies, DARIAH-DE aims to ensure the long-term preservation and future use of research data. DARIAH-DE also supports and advises researchers as well as research projects in planning and accomplishing humanities research initiatives within a digital environment. Effective ways of handling digital resources, concepts, and methods in the digital humanities must be introduced into training and instruction for humanities researchers at all educational and career levels. In close consultation with disciplinary communities, DARIAH-DE coordinates and (where necessary) further develops existing study and training courses. Moreover, DARIAH-DE is developing individual qualification modules, such as international workshops for experts dealing with specific themes. This infrastructure will enable researchers to carry out research in an increasingly digital environment, across disciplines and institutions in collaborative ways and towards sustainable results.

For the first two and a half years of the DARIAH-DE project, the computing centers and software partners participating in the consortium have laid the groundwork for the establishment of a sustainable technical infrastructure based on the requirements and needs of researchers in the arts and humanities. Although a productive research infrastructure now exists, there are still some requirements to be addressed in the coming years, together with representatives from the computing centers, the information specialists and the software developers, and partners from the various disciplines in the humanities. As a research-driven infrastructure project, DARIAH-DE

demonstrates that both research projects as well as national and international collaborations in particular need sustainable research infrastructures. Currently there are more than 500 researchers from a wide spectrum of disciplines in the humanities and about 35 projects using the options and services provided by Dariah-DE, such as the Confluence Wiki system, the Developer Portal, and resources such as VMs, storage, and discipline-specific services. To enhance the stability and usability of research data, Dariah-DE is building a research data repository for the arts and humanities. This repository supports both the storage of data in the creation process as well as the long-term preservation of research data.

Dariah-DE is a reliable research infrastructure in Germany for the arts and humanities in the field of research, teaching and research data. It is, furthermore, the German national contribution to the European research infrastructure Dariah-EU within the framework of ESFRI. External research projects from the arts and humanities and service- or tool providers are invited to cooperate with Dariah-DE to provide their services and tools in the Dariah-DE research infrastructure.

The aim of this Dariah-DE poster presentation at the DH 2014 in Lausanne is not only to present a poster describing Dariah-DE; it will also include a multi-media presentation of different kind of demonstrators in the fields of research, teaching, research data, and technical infrastructure. The conference participants will be presented with a broad variety of (re-usable) specific and interdisciplinary tools and services from the Dariah-DE research infrastructure. The overall objective of the multimedia presentation is to show a variety of Dariah-DE tools and services ready for researchers to use. Examples include DIGIVOY, a bundle of text analysis tools, the Dariah-DE Geo-Browser, a tool to analyze time-spatial relations, the Collection Registry and a generic search engine. We will also present tools for collaborative work as well as our service-oriented web portal (de.dariah.eu), based on Liferay, which allows for the embedding of externally-developed services and tools.

More information about the German project Dariah-DE and the European project Dariah-EU can be found here: <https://de.dariah.eu/> (German) and <https://www.dariah.eu> (English).

Contact address:

dariah-sub@sub.uni-goettingen.de Website: <https://de.dariah.eu/>

The Text portal: An online resource providing medieval literature for students and their teachers

Schneider, Gerlinde

gerlinde.schneider@uni-graz.at

Centre for Information Modelling - Austrian Centre for Digital Humanities, University of Graz

Schwinghammer, Ylva

ylva.schwinghammer@uni-graz.at

Department of German Studies, University of Graz

Teaching material on medieval German literature adequate to present-day educational standards is insufficiently available. Additionally, the subject tends to be marginalized in current curricula for teacher training. Consequently, schoolteachers often lack the impetus and time to prepare and arrange medieval texts for their students.

The objective of the publicly funded "Sparkling Science" project "Arbeitskoffer zu den Steirischen Literaturpfaden des Mittelalters" (Toolbox for the Styrian literature paths of the Middle Ages) is to develop an openly accessible, virtual didactic environment (the "Text portal") which enables teachers to address the topic of medieval literature in class and familiarize their students with older German texts. For that purpose a corpus comprising regional medieval literature is made available on the Text portal by providing the transcription,

translation and facsimiles of the respective texts. Furthermore, glossarial information and additional teaching material adapted to the requirements of students on secondary educational level are featured.

This poster introduces the online resource outlined above which is developed at the University of Graz in a collaboration between the Department of German Studies and the Centre for Information Modelling. The main focus is set on the following two aspects which are fundamental for the realization, progress and sustainability of the project:

1. The inclusion of prospective users

The project primarily targets two disparate groups: Teachers in secondary education on the one hand and their students on the other hand. In order to provide an efficient and effective resource, the requirements of both user groups were considered the guiding principles for the development of the portal and its underlying data, from the very beginning.

The initiators of the project are scholars of German Studies involved in the education of both teachers and younger students. In addition to their individual views on the issue, the requirements to this resource were elaborated in the context of university courses for teacher training in concert with actual teachers and teacher trainees.

To investigate the demands of the younger students, a multilayered empirical analysis was conducted amongst several hundred pupils to survey the acceptance and comprehension of Middle High German and Early New High German texts. The outcome of this survey directly influenced later decisions concerning the selection and edition of the texts provided.

Taking account of the diversity of the prospective user groups, an intuitive, user-friendly and dependable interface will be realized. With this goal in mind, the development of the portal is subject to continuous evaluation within the project team. Further evaluations and test runs in a real-life environment with students and teachers are scheduled as the project progresses.

2. The involvement of Digital Humanities competence

The literary texts provided on the text portal are chosen with regards to their supraregional importance or their interest within the medieval literary canon, with the majority not yet being available in a standardized digital format. Enrichment and annotation methods common in Digital Humanities and the usage of trusted digital infrastructures serve to produce digital resources in this particular project which are later reusable in other contexts and sustainably preserved.

All textual sources used within the project are encoded conforming to the guidelines of TEI P5, annotated and documented in a way that makes an adaption to further editorial works or other projects as unproblematic as possible. The TEI representations as well as the pictorial representations and the metadata of the original sources are openly available via the FEDORA based asset management system GAMS.

References

Hockey, S. (2012). Digital Humanities in the Age of the Internet: Reaching Out to Other Communities. In Deegan, M. and McCarty, W. (eds), *Collaborative Research in the Digital Humanities*. Farnham: Ashgate, pp. 81-92.

Kümper, H. (2011). Mittelalter und Mittelalterunterricht im neumedialen Zeitalter. In Kümper, H. (ed.), *eLearning & Mediävistik*. Frankfurt am Main: Peter Lang, pp. 7-65.

Schmidt, S. (2004). Neue Medien und alte Themen: Internet & Co als Vermittlungshilfen in der Schule, im Museum und an der Universität. In van Eickels, K., Weichselbaumer, R. and Bennewitz, I. (eds), *Mediaevistik und Neue Medien*. Ostfildern: Thorbecke, pp. 243-259.

TEI Consortium (2013). *TEI P5: Guidelines for Electronic Text Encoding and Interchange, Version 2.5.0*. Virginia: Text Encoding Initiative Consortium Charlottesville.

Warwick, C. (2012). Studying Users in Digital Humanities. In Warwick, C., Terras, M. and Nyhan, J. (eds), *Digital Humanities in Practice*. London: Facet Publishing, pp. 1-22.

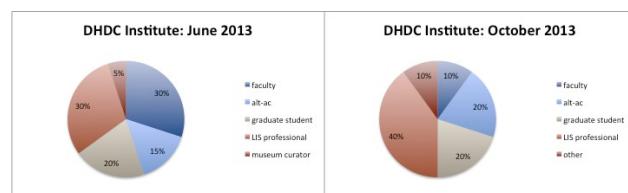


Fig. 1: Participant demographics from DHDC Institutes

Digital Humanities Data Curation Institutes: Challenges and Preliminary Findings

Senseney, Megan

mfsense2@illinois.edu
University of Illinois

Muñoz, Trevor

tmunoz@umd.edu
University of Maryland

Flanders, Julia

j.flanders@neu.edu
Northeastern University

Fenlon, Ali

fenlon2@illinois.edu
University of Illinois

1. Introduction

The growth of digital humanities research has made the curation of DH research data a priority for humanities scholars and their institutions. Data curation “addresses the challenges of maintaining digital information produced in the course of research in a manner that preserves its meaning and usefulness as potential input for future research”¹. More fully integrating data curation into digital research involves fluency with topics such as publication and information sharing practices, descriptive standards, metadata formats, and the technical characteristics of digital data. This poster presents lessons learned from a series of workshops on digital humanities data curation conducted in June and October 2013 with funding from the National Endowment for the Humanities Institutes for Advanced Topics in the Digital Humanities program.

2. Project Background

The Digital Humanities Data Curation (DHDC) institute series is a collaborative initiative led by the Maryland Institute for Technology in the Humanities (MITH) in cooperation with the Women Writers Project (WWP) at Northeastern University and the Center for Informatics Research in Science and Scholarship (CIRSS) at the University of Illinois’ Graduate School of Library and Information Science (GSLIS). Institutes are designed to offer humanities scholars with all levels of expertise a grounding in data curation practices and problems.

3. Workshop Participants and Community Response

Two of three institutes have been conducted, first in June 2013 at GSLIS and again in October 2013 at MITH. The team received 111 and 136 applications for the first and second institutes, which were designed to accommodate 20 participants each; acceptance rates ranged from 15% to 18%. For an overview of participant demographics across institutes, see Fig. 1.

4. Participant Feedback and Lessons Learned

To make the curriculum responsive to participant needs, the project team conducted evaluation surveys following each workshop and analyzed materials generated from workshop activities, note-taking, social media interactions, and direct feedback. Responses thus far have expressed enthusiasm for instructors and the workshop’s overall framework, while also recommending the addition of more hands-on activities and a greater focus on tools, metadata, and infrastructure issues. Participants have consistently ranked the following topics as very valuable for digital humanists: conceptual frameworks for data curation, understanding the nature of digital objects, types of metadata, collections as curation, and data and the law. To address participant feedback following the first institute, the project team revised the initial curriculum by increasing the number of group exercises, adding two lectures on metadata, and including a hands-on session on item deposit and retrieval in Islandora. The schedule for each institute is available at <http://www.dhcuration.org/institute/schedule/>.

5. Significance of Digital Humanities for Data Curation

One key outcome of the institutes thus far has been a set of emerging insights into the special challenges of data curation in a digital humanities context, which was an important goal of the original proposal. These insights have been particularly evident in several key areas of the discussion. First, each institute has featured a discussion of roles that has revealed the diversity of job descriptions and professional identities of data curators in digital humanities. Second, the featured case studies have demonstrated the challenging nature of digital humanities data with respect to format, anticipated future usage, and methodological texture requiring capture and documentation. And finally, participant questions have shown a need for resources to guide data curators in working with specifically digital humanities data: for example, crowd-sourced transcriptions, standoff annotation, and data in experimental formats.

6. Conclusion

The overwhelming response to DHDC’s calls for applications indicates the need to sustainably conduct data curation training for digital humanists at a larger scale. While, slide sets, notes, and resources lists are currently available online through the project’s GitHub wiki (<https://github.com/digital-humanities-data-curation/dhdc-workshop/wiki>), project materials will soon be revised for broader impact and integrated into the affiliated DHCurator Guide (<http://guide.dhcuration.org/>), a community resource that extends beyond the project in promoting a community of scholars focused on discipline-specific curation practices and skills.

References

1. Muñoz, T., & Renear, A.H. (2011). *Issues in humanities data curation*. Discussion paper circulated at the Palo Alto Summit on Humanities Data Curation, Stanford, CA, June 23, 2011. Available at <hdl.handle.net/2142/30852> and <cirssweb.lis.illinois.edu/paloalto/whitepaper/premeeting>

Adams Family Legacy: Visualizing the World of an American Presidential Family

Sikes, Sara

ssikes@masshist.org
Massachusetts Historical Society

Christian-Lamb, Caitlin

cachristianlamb@davidson.edu
Davidson College

Spanning the years 1735 to 1889, the Adams Timeline (www.masshist.org/adams/timeline) is a searchable collection of key events and happenings in the lives of 2nd U.S. President John Adams, First Lady Abigail Adams and three succeeding generations of their immediate family. Members of the Adams family were deeply involved a tumultuous era of American history and were keen observers of national and domestic politics, as well as the daily activities on their beloved family farm. The collection of Adams Family Papers at the Massachusetts Historical Society is the most comprehensive and historically complete family collection held by any American cultural institution. While forming the basis of numerous digital and analog resources, this vast body of material lacked a coherent summation of major personalities and collection highlights.

The creation of the Adams timeline achieves the dual results of a streamlined presentation of historical data and fulfillment of a need in the research community. A diverse audience ranging from published scholars to schoolchildren land directly on this resource when searching for biographical information on the Adams family and access an interactive organization of key data points. Residing on the website of the Massachusetts Historical Society, the timeline acts as a portal for locating different types of Adams family information held by the Society. This was an initially unforeseen benefit of creating the timeline, but the addition of hyperlinks to transcriptions and images of original documents allowed for ready access to related materials, including collections of letters and transcriptions, images of diary entries and annotated documents from our Adams Papers Digital Edition.

This timeline was built as a customized adaptation of the SIMILE timeline module (simile-widgets.org/timeline), part of a suite of open-source data visualization widgets originally developed at the Massachusetts Institute of Technology. Designed to handle specific dates, time spans, events, images and links, the timeline is rendered from data in an underlying XML file. Each individual is also encoded with a unique identifier, allowing for filtering of events relevant to a certain person and the creation of a focused timeline for an individual rather than the whole family.

While an earlier version of the timeline was displayed only as a static table, this newly created web tool visualizes temporal information and allows for the analysis of the intersection and overlapping of interrelated events. A well-designed data visualization allows users to quickly spot patterns, trends, clusters, gaps and outliers and fulfills Maureen Stone's definition of information visualization, as "the creation of graphical representations of data that harness the pattern-recognition skills of the human visual system".¹ The Adams timeline now allows for users to make ready connections through time, understand relations between events and within context and quickly scan a dataset in ways that were not possible within a static table. As Joseph Priestley noted in his 1764 publication of a chronological chart representing historical figures, "the thin and void places in the chart are, in fact, not less instructive than the most crowded."² Thus a gap in a timeline may be just as meaningful as an area of high activity and an opportunity for exploration of the underlying causes of such a void.

The aim of this poster presentation is two-fold: to explore the process of creating an XML-based timeline with SIMILE widget; and to demonstrate a possible mode of delivery for the vast stores of information held within the walls of public

and private cultural institutions. We will offer an interactive demonstration of our customized adaptation of the SIMILE widget timeline module and solicit input from DH2014 attendees on the methodology and creation of complementary interactive tools. We also seek to explore other possibilities for delivering the wealth of historical data in our collection, currently housed in binders or spreadsheets. We envision this timeline as a first step in designing additional tools, such as a map of Adams family residences or a visualization of the correspondence network of a family deeply connected to early American history.

References

1. **Maureen Stone** (2009), *Information Visualization: Challenge for the Humanities*, Working Together or Apart: Promoting the Next Generation of Digital Scholarship, 145:43–56 (March 2009). Available at www.clir.org/pubs/resources/promoting-digital-scholarship-ii-clir-neh/stone11_11.pdf.
2. **Joseph Priestley** (1764), *A Description of a Chart of Biography*, Warrington, Eng.

Empowering Student Digital Scholarship: CLASS Program as a model for digital humanities scholarship in the Liberal Arts

Simons, Janet Thomas

Hamilton College, United States of America

Nieves, Angel David

Hamilton College, United States of America

Grimaldi, Kerri

Hamilton College, United States of America

Proposal

Culture, Liberal Arts, and Society Scholars (CLASS) is an undergraduate research and fellowship program in the digital humanities awarded to student scholars at Hamilton College's Digital Humanities Initiative (DHi). Basic literacies for the digital age are critical skills sets for students entering the professional world in the twenty-first century. The Digital Humanities Initiative provides new opportunities for students in the humanities to become fully engaged citizens in this ongoing digital revolution.

CLASS is based on three-broad areas of scholarly inquiry and their intersection with new and emerging digital technologies: 1) Culture, 2) Liberal Arts, and 3) Society. CLASS provides a unique partnership between departments, programs, and units across the liberal arts and humanities at Hamilton in partnership with the College's Career Center. It begins with course connections in our Cinema and New Media Studies (CNMS) program but then removes the confines of the semester to promote deep understanding of digital humanities research within a specific field of interest. In these experiences, students and their faculty advisor become part of a collaborative working team of experts in DHi.

CLASS provides students with skills training in digital literacies through intensive research and scholarship coupled with two unique internship experiences. In the summer between sophomore and junior years CLASS offers undergraduate students an intensive professional development experience and provides a comprehensive overview of work in their respective field of interest. In the summer immediately after their junior year students enter their second internship off campus leading to employment and/or graduate study as a result of the eighteen-month program. Assistance with job placement, in a professional field, based on their CLASS internship placement, and/or graduate studies occurs in their final year at Hamilton.

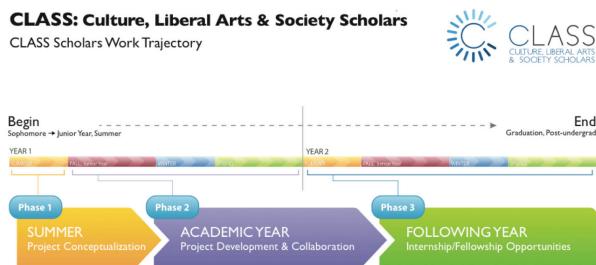


Fig. 1: Fig. 1. CLASS Program 18-month Structure

The coursework for the program begins in the fall of their sophomore year with CNMS 120 or 125 (Fig 1). In the spring semester students can enroll in courses offered in the CNMS minor. Students enroll in either CNMS 200W/Introduction to Digital Humanities or CNMS300/Interdisciplinary Research Methods that provide experiences writing grant proposals in the digital humanities.

The goals for CLASS include:

- Collaboration with potential faculty and/or staff mentors to define and develop an interdisciplinary project
- Writing a research proposal for their projects.
- DHi committee reviews the proposals and recommend possible award opportunities.
- Students begin work over a 10-week period in the summer, mid-June to late August.

A two-week intensive training program takes place in June of the first summer. Students survey mature digital humanities projects, participate in discussions of digital humanities readings, interact with invited speakers brought into the program, and explore technologies related to their research project goals. During the academic year following the first summer, students work with their mentor between 4-6 hours a week on their collaborative research project. In the summer between junior and senior years, CLASS offers undergraduate students an intensive professional development experience and provides a comprehensive overview of work in new digital technologies.

CLASS will:

- Develop understanding of digital humanities methods
- Develop technological expertise for careers and digital scholarship

Through their participation in an undergraduate research project, students will be able to:

- Develop a research question, problem, or design;
- Apply basic principles and knowledge found in the literature related to the research question;
- Develop a research proposal to address or resolve a specific research question or problem;
- Apply and evaluate interdisciplinary methodologies throughout the project;
- Collect, interpret, and critique data in order to resolve a research question or evaluate a design;
- Utilize digital skills (TEI, digital collection development, media object creation, geospatial visualization, etc.) necessary for robust digital scholarship in the humanities
- Communicate research findings through oral presentations and digital publications.

Students in research collaborations with Faculty and members of the DHi, develop deep understanding of a specific long-term research agenda. They are expected to conduct collaborative investigation of a specific aspect of the research that is of great interest to them and integrate digital humanities research methods in their process. Deliverables include public presentation and/or publication at milestones in this process.

Accomplishments:

Cohort 2011

Sarah Bither and Melissa Yang worked with Professor Kyoko Omori to develop an understanding of Benshi performance

art for contemporary audiences. The outcome of this work is a website, <http://courses.hamilton.edu/dhi-class-1/sarah> with components that will ultimately be incorporated into Professor Omori's *Japanese Comparative Literature Archive*. Bither continued her study of Japanese culture and language by going abroad to Japan in the spring and summer of 2012. Randall Telfer worked with Professor Thomas Wilson to explore Confucian rituals and connections to contemporary religious practices in China. The outcomes of this work were additional edits to two of Professor Wilson's websites http://academics.hamilton.edu/asian_studies/home/asc_test/index.html and The Cult of Confucius website: http://academics.hamilton.edu/asian_studies/home/coc_test/index.html. Brynna Tomassone worked with Professor Angel David Nieves to explore cultural connections between South Africa and the United States during the time period leading up to the events in 1976 Soweto. The outcome of this work is a series of book chapters currently in forthcoming publications including *The Heritage of Iconic Planned Communities: The Challenges of Change* (University of Pennsylvania Press, 2014). Tomassone is now a Ph.D. student in Hispanic/Spanish Studies at Syracuse University.

Cohort 2012

Maxwell Lopez ('14) mentored by Professor Nathan Goodale. Continuing aspects of research on the history and culture of the Sinixt Nation in British Columbia, Lopez proposed to work for the 2012 -2013 year creating, "an accurate three dimensional digital representation of the British Columbia site along with some models of artifacts excavated" using the Unity game engine. He believes the project will create a "new way for people to experience and interact with history" and have great capacity for connecting with research with the public. By the end of the two- week CLASS session in June 2012, Lopez had already made several models in Blender (a 3D modeling software) and site maps in Unity (a virtual world platform for models to reside). The following is a screenshot of his initial construction of a pit house in Blender. Please see the folder on CLASS 2012 for screenshots of his work to date and his complete proposal.



Fig. 2: Figure 2. Sinixt Ritual Pit House model (created in Blender).

Lopez continued work with Professor Nathan Goodale in GIS mapping and archeological data collection for continued development of Sinixt Ritual Pit (Fig. 2) houses at an archeology Field School in British Columbia summer 2013. Lopez has presented aspects of his work with Professor Goodale at several forums on campus in Fall 2012 and also at the 2013 Re:Humanities symposium April, 5, 2013.

Continuing aspects of O'Neill's development of *Beloved Witness: Agha Shahid Ali Archive*, Ujjwal Pradham ('14) will explore the use of text analysis tools and TEI in developing aspects of the *Agha Shahid Ali Archive*. Pradham is also interested in establishing a connection between the archive and current communities of interest in Kashmir.

Pradham explored the uses of social software to make connections between contemporary Kashmir communities and the developing *Beloved Witness* archive. The following is a

screen shot of the Voyant Tools (Fig. 3) text analysis Ujjwal did to compare theme words (home, waiting, never, Spring, Kashmir) in multiple manuscripts over the development of a poem written by Ali.



Fig. 3: Figure 3. Voyant Tools.

Cohort 2013

Working with Religious Studies Professor Abhishek Amar on aspects of his Sacred Centers in India Project, students Kenneth Ratliff ('16) and Alex Gioia ('14), embarked on a study of Indian sacred centers -- Buddhist Bodhgaya and Hindu Gaya. The students expanded their understanding of the Indian sacred cities of Gaya and Bodhgaya. They assisted Professor Amar in organizing his research data for these two cities (images, videos, and GPS coordinates) into the metadata schema for his digital research archive. This work of organizing and processing the over 418 data objects from Gaya into survey forms conducive to further analysis is necessary for long term sustainability of the digital archive. It is also the first step in the creation of interactive models of important artifacts and their locations within these religious sites. Appendix "A" is an example of one of the individual data survey forms used in this project and is based on those used by the K.P. Jayaswal Research Institute, with whom Professor Amar collaborates. Ratliff and Gioia have already begun creating an interactive two-dimensional line map of the Vishnupada complex (Fig 4).

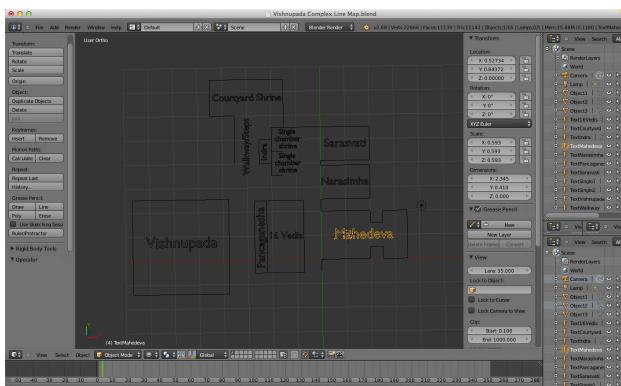


Fig. 4: Figure 4. Interactive line map of Vishnupada Complex with Mahadeva site highlighted (created in Blender).

This interactive map will link to images of the sites (Fig. 5) and 3D models of the artifacts in situ (Fig. 6). They hope that these virtual 3D and geographically correct models will foster greater interest in these religious sites due to the accessibility and interactivity of the maps, photographs, models, and videos.

Ultimately, the plans are to place the models that they produce into an online viewing space, developed from a game engine (Unity), that will make the models easily viewable and web accessible to the public.

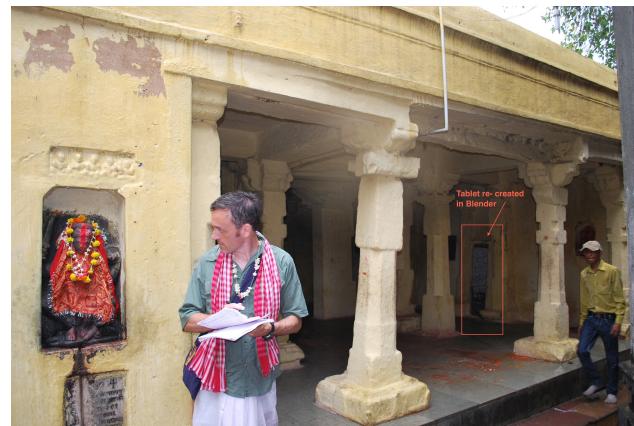


Fig. 5: Figure 5. Image of Mahadeva site and Tablet artifact on far wall.

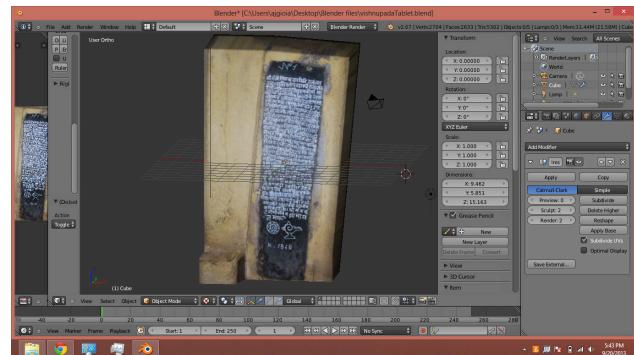


Fig. 6: Figure 6. Image of Mahadeva Tablet being modeled in Blender (Free Download at Blender.org).

Working with Patricia O'Neill on aspects of her Beloved Witness archive, Kerri Grimaldi ('16) examined the significance of Emily Dickinson's poetry to Agha Shahid Ali, a poet from Kashmir, whose work is the focus of the archive. Grimaldi's project traces the depth of Emily Dickinson's influence in Shahid's poem, "A Nostalgist's Map of America," by placing Shahid's poem side-by-side with Dickinson's "A Route of Evanescence" in four stages of analysis, each increasing in level of explication. By analyzing Shahid's poem, it is possible to read Dickinson's in a completely different light, while also witnessing the resonating power of her poetry. Grimaldi has started to create a website to present her analysis of the relationship between the work of Shahid and Dickinson. Ultimately, this website will show not only that Dickinson influenced Shahid's work, but that his work responded to and interpreted hers, such that their works are in conversation with each other. Please review the current status of this project, including the descriptive first layer of the website, the storyboards exploring intertextuality in the second layer and third layers and a draft of Grimaldi's own creative video interpretation of the two poems in conversation. This project was submitted in September 2013 to the Dickinson Electronic Archives 2.0: CALL FOR PROPOSALS for volume 3 -- *Emily Dickinson's Reading Culture* to be published in 2014. Part One of Kerri's website can be found at <http://dhinitiative.org/demos/grimaldi/>

Challenges

Initial Challenges for DH in developing CLASS included answering, how do we publish in the digital Humanities? Much of collaborative research includes the use of copyrighted material and/or faculty research that is still in early development. These characteristics of the work in CLASS required that we reconsider the amount and type of information (public or not) conveyed to illustrate the progress of the students. We achieved our goal of facilitating discussion with other scholars in the field by including experts from across disciplines in the two-week intensive training program and through the natural association of collaborators on the faculty

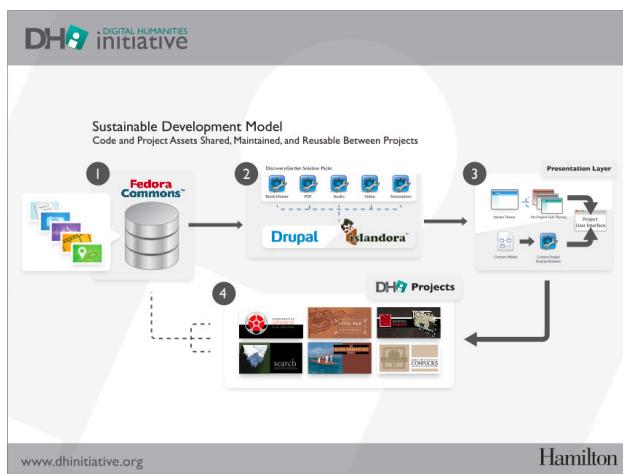
research projects. CLASS scholars biographies and research descriptions are announced on the DHi website. CLASS scholars give project prospectus presentations and/or example research projects at the end of their two-week training program in the first summer. These presentations and examples are given to an invited audience for feedback on the projects. Ultimately, each student presents or publishes their work off-campus. Several students have presented at Re:Humanities.

CLASS Program Summary and Future Plans

Students in research collaborations with Faculty and members of the DHi, have been successful in developing deep understanding of a specific long-term research agenda. DHi provides the immersive experiences and ongoing continuity with faculty research necessary to support this engagement. Most of our CLASS scholars have conducted collaborative investigation of a specific aspect of a long-term research agenda, determined their specific interests and contributions to that agenda, and publicly presented their scholarship "in progress" at Hamilton events and professional conferences.

Our future plans are to continue the CLASS program and to collaborate on its development with other liberal arts schools. Several schools have asked us about building similar models at their schools. This would require thinking more about scaling-up the program. One option we are considering is a form of "Summer Institute" for undergraduates in which we bring them together with their mentors and a larger DH community for a two-week program.

Appendix A: Proof of Concept for a Sustainable Digital Humanities Faculty Collection Infrastructure in the Liberal Arts.



DHi's technology infrastructure and research support model is designed to be sustainable. That is, our approach will reduce the need for regular revamping of static faculty research web pages by creating infrastructure and processes that maintain research outcomes as "living" web presences accessible for faculty and student collaborative scholarship over time. To this end we researched best practices in digital collection development and preservation in collaboration with members of our library and decided to develop an institutional warehouse (repository) for digital collections (Fedora Commons). Fedora was chosen for its scalability and ability to be extremely flexible in the way objects can be accessed. Fedora has built-in flexibility to allow creation and maintenance of relationships among objects and across digital collections over time.

After researching open source collaborative tools to interface with collections in Fedora Commons we decided to make use of Islandora. Islandora can be used to create customized themes for faculty collections and projects. Our DHi Collection Development Team is working with the Islandora and Fedora Commons consultants (at Discovery Garden to create our digital scholarship infrastructure. By using experts to help with development we are making efficient use of the Mellon Grant to move this complex project forward.

2013 Appendix B: KPJR Form Vishnupada Complex Mahadeva Temple

DOCUMENTATION SHEET OF BUILD HERITAGE/SITE
N.M.M.A., ARCHAEOLOGICAL SURVEY OF INDIA
COMPILED AT K.P. JAYASWAL RESEARCH INSTITUTE,
PATNA

Sl.No.	Documentation Parameters	
	State/Dist./Block	Gaya
1	Name of the monument/built heritage/site	Mahadeva Temple
2	Date/Period	Early Medieval/Medieval?
3	Location	To east and north of 16 Vedis/Padas in Vishnupada Complex
	Geo-coordinate	
4	Approach	East of the Vishnupada Temple, in the Vishnupada Complex
	Airport	Gaya
	Railway Station	Gaya
	Bus Stand	Gaya
5	Topographical features	Slope of the Mundaprishta hill on the western bank of the Phalgu
6	Brief History	Temple seems to have origins in early medieval or medieval period. The exact date of the construction of the temple is difficult to determine because of lack of historical sources. It has images and inscriptions, but they may have been moved.
7	Local tradition associated with building/structure/site	Gaya Shraddha, place of Pinda-dana as well as Darshana
8	Architectural style	Inner sanctum, which has a Linga, and there are 20 pillars, which constitute the Mandapa
9	Description of the building/structure/site	Mahadeva temple: Inner sanctum with a Shiva Linga and a twenty pillared Mandapa. Narasimha Temple (east of Mahadeva): Small rock temple Sarasvati Temple (east of Narasimha): Small rock temple Facing Narasimha are two single chamber shrines. All five are treated as one unit in the Vishnupada complex.

10	Building/Structural material and other	Stone and brick; pillars are stone
11	Usage(s)	Active worship, Darshana
12	Ownership	Same as Vishnupada Main Temple
13	Protection status	Good
14	Present condition	Maintained
15	Conservation assessment	Alright
16	Photographs	See attached
17	Plan/elevation, if available	
18	Published references	
19	General Remarks	There is an inscription
20	Name and address of compiler with date elements used	Abhishek Singh Amar Matthew Sayers

Images:

Mahadeva Temple (129):

By opening

1. Vishnu, 16" (122)
2. Camunda, 16" (120, 121)

– There's no logic to the temple - they have plastered images all over the place; normally, you would not see Vishnu and Camunda next to each other, but in this case we do, in a disorganized fashion. For instance, they are in the same niche, but Camunda is plastered higher than Vishnu, which speaks to the disorganization of the collection process.

South wall

3. Ganesha, 16" (122)
4. Inscription (123-125)

– Appears to be painted more recently.

5. Eroded Uma-Maheshvara, 14" (126)
- Inner sanctum (127)

6. Huge Shiva Linga (127)

North wall

7. Uma-Maheshvara, 16" (128)

By opening, on the north, west-facing niche; left

8. Durga, 36" (130)

9. New inscription (130)

Outside, left

10. Dasavatara (131)

Reading Between the Lines: Image-to-Segment Relationship Development and Analysis

Smith, Dustin*dusin.smith@utexas.edu*

The University of Texas at Austin, United States of America

Karadkar, Unmil*unmil@ischool.utexas.edu*

The University of Texas at Austin, United States of America

Galloway, Pat*galloway@ischool.utexas.edu*,

The University of Texas at Austin, United States of America

Davis, King*king.davis@austin.utexas.edu*

The University of Texas at Austin, United States of America

1. Introduction

Crowdsourced transcription has been adopted successfully by institutions large and small in order to unlock cultural heritage data in handwritten documents. In the case of mental health documents, however, public transcription is not an option as the act of transcription involves reading them, which has implications for privacy of patients, doctors, and employees alike, potentially resulting in social repercussions to family and descendants. We are exploring mechanisms to crowdsource the transcription of such privacy-sensitive documents while maintaining the anonymity of individuals named in handwritten records by constraining the context of such mention as well as by exploiting other characteristics of documents. We report our promising initial results and describe our approach for generating structural metadata to identify multi-page units within large registers to improve the granularity of access.

1.1. Collection

Now called the Central State Hospital (CSH) [1], the first mental health institution for African-Americans in the USA was founded in 1870 near Petersburg, VA. The meticulous records maintained by the custodians of this historically significant institution have now stored as tiff files at 400 dpi resolution in a folder structure that reflects minimal structural information. The documents include hospital administrative and medical records of all stripes. The early records are handwritten and must be transcribed before these can be analyzed to observe patterns in administrative practices, as well as patient care. While the data set contains several types of handwritten cursive documents, including patient records, we are basing the development on board meeting minutes in order to minimize the risks in case of accidental exposure of these records.

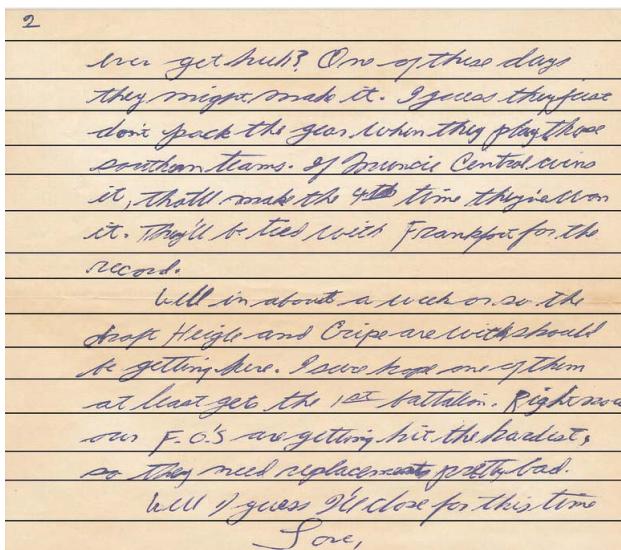


Fig. 1: Cursive document with line breaks using adaptive thresholds.

1.2. Prior work

There is a vast body of literature on off-line handwriting recognition, focusing on methods for automatic character identification and transcription into machine-readable text [2]. Two examples of character recognition work are Tomai et al., [3] who provide a framework for mapping words in a transcript to a word image in a document and Guillevic et al., [4] who provide a character recognition approach to unconstrained, small-lexicon cursive handwriting. However, our corpus is heterogeneous, authored by many individuals, and uses a broad vocabulary.

2. Approach

We are using the Gamera libraries to segment documents for identifying individual words, in an effort to minimize the context available to potential transcribers, much like captchas are used to transcribe old, hard to OCR documents. As each document is different, and the text across lines overlaps in various ways, ruling out the use of bounding box-based methods. We are using flexible, self-adjusting thresholds to detect lines. Fig. 1 shows the identified line breaks, with some breaks going through ascenders or descenders within characters in a line based on thresholds detected using document histograms. Additionally, we are using X-Y histogram profiles to identify and mark meeting minutes that span multiple pages (but begin on a fresh page) to improve the access granularity for these documents. Fig. 2 shows the current document structure using solid lines (pages within a register) and the intermediate minute structure superimposed by dashed lines.

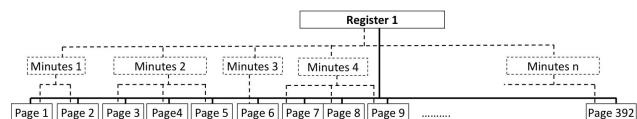


Fig. 2. Generation of multi-page document structure

Fig. 2: Generation of multi-page document structure

3. Discussion and Future Work

References

By employing a top-down, histogram-based approach to line and word recognition, our methods are adaptable to a large body of handwritten documents that space text at varying distances. In addition, the histograms also enable the generation of structural metadata that improves the access granularity.

The next step is to identify words within each line, where the threshold-based approach will help us locate spaces between inclined words. Transcribing at the word-level will enable us to expose varying levels of documents contexts to potential transcribers, depending upon a computation assessment of privacy sensitivity of content (for example, very short words are less likely to contain individual information). Prevention of identity disclosure in these records is critical due to the stigma associated with treatment for mental health issues. The methods we develop for transcribing handwritten documents will also be applicable to other privacy-sensitive historical records, most immediately, those of similar mental health institutions that followed the CSH.

4. References

The Central State Hospital. <http://www.csh.dhhs.virginia.gov/>

Cattoni, R., Coianiz, T., Messelodi S., and Modena ,C. M. (1998). 'Geometric Layout Analysis Techniques for Document Image Understanding: a Review' January 1998

Tomai, C. I., Zhang, B. and Govindaraju V. (2002).

Transcript Mapping for Historical Handwritten Document Images. Proceedings of the Eighth International Workshop on Frontiers in Handwriting Recognition.

Guillevic, D. and Suen, C. (1995). Cursive Script Recognition applied to the Processing of Bank Cheques.

Taking a Global Perspective on the Skills and Competencies Important to Digital Scholarship

Spiro, Lisa

Digital Scholarship Services, Fondren Library, Rice University, United States of America

Cawthorne, Jon

Florida State University Libraries

Lewis, Vivian

McMaster University Library

Wang, Xuemao

University of Cincinnati Libraries

Despite the significance of digital technologies to contemporary culture, scholars across the mainstream academy have been slow to publish “discipline-based scholarship produced with digital tools and presented in digital form” (Ayers, 2013). In part, this dearth results from many researchers lacking (or lacking access to) required skills (a learned capability, such as programming expertise) and competencies (a more abstract ability in a particular domain, such as transdisciplinary collaboration). While some centers and programs excel at supporting digital scholarship, many institutions are trying to determine what skills are needed to produce such work and how to cultivate them (Sehat and Farr, 2009). Our benchmarking study thus examines how leading digital scholarship (DS) programs around the world define and nurture skills and competencies. With the support of the Mellon Foundation, we are conducting site visits and gathering data at four locations in the United States, one in Canada, two in Europe, and three in the BRICS countries (Brazil, Russia, India, China and South Africa). As a result of this research, we hope to help institutions create strategies for supporting digital scholarship, inform the development of educational and professional development programs, and contribute a global perspective to the ongoing discussion about DS training. Although this study is necessarily constrained to approximately ten sites, we hope that it provides the foundation for further work.

Through benchmarking, we aim to understand the key workforce-related factors instrumental to a center’s, department’s or program’s success. In order to consider a range of methodological approaches and foster cross-disciplinary inquiry, we define digital scholarship broadly as creating, producing, analyzing, and/or disseminating scholarship using new technologies, with an emphasis on digital and computational techniques. Digital scholarship includes both creating and analyzing “born digital” content and bringing new meaning to digitized content, such as through textual analysis. Our study looks at digital scholarship not only in the humanities, but also in the social sciences. We are investigating both physical centers and distributed DS services and activities, which are often, but not always, located in a university library. In addition, we are examining to what extent cultural, institutional, disciplinary and national contexts influence what skills are important.

This poster will serve as an interim report on the study, outlining its goals, methods, and initial high-level observations. In selecting sites, we are considering criteria such as what research services they offer, staff expertise, numbers of different kinds of staff, reputation, and record of innovative, successful projects, as well as the program’s involvement with significant professional development and/or educational programs. We aim for geographical, cultural and disciplinary diversity. During our site visits, we are conducting semi-structured interviews with key faculty, graduate students, research staff, and administrators, asking questions such as what skills matter most to their work and how ideally to develop them. We are also gathering benchmarking data such as the skills of staff, the center’s support for professional development, and its approaches to developing digital scholarship.

By the time of the conference, we expect to have completed at least six of our site visits, including to locations in the United States, Canada, China, and Europe. Thus we will share general patterns emerging from the study, including the preliminary core skills and competencies that we have identified. We will also explore how centers develop these skills, such as by providing dedicated time for exploratory research and development work, hosting visiting scholars, enabling staff to

teach, supporting training and conference travel, and fostering a “learning culture.” We will discuss the challenges in conducting the study, including identifying “best in class” digital scholarship programs in a global context, ensuring diversity in what centers and programs are represented, coordinating visits, and deciding what kind of data is important to collect and how best to capture that data. Through our interactive poster session, we will solicit feedback on our work to date and identify opportunities to expand our research.

References

- Ayers, Edward L.** (2013) *“Does digital scholarship have a future?”* EDUCAUSE Review, vol. 48, no. 4 (July/August 2013)
- Sehat, Connie Moon, and Erika Farr.** (2009) *The Future of Digital Scholarship: Preparation, Training, Curricula. Report of a Colloquium on Education in Digital Scholarship, April 17–18, 2009.* Washington, DC: Council on Library and Information Resources, 2009. www.clir.org/pubs/archives/SehatFarr2009.pdf.
- Zorich, Diane** (2008). *A Survey of Digital Humanities Centers in the United States.* Washington, DC: Council on Library and Information Resources. <http://www.clir.org/pubs/abstract/pub143abst.html>.

Converting Medieval Documents into a Searchable Database

Sporleder, Caroline

Trier University

Fertmann, Susanne

Saarland University

Krones, Tim

Saarland University

Kolatzek, Robert

Saarländische Universitäts- und Landesbibliothek, Verlag universaar

Teufel, Isolde

Freiburg University Library

1 Introduction

More and more historical documents are digitised. In their raw form, however, digitised sources are usually unstructured texts, on which only keyword search is possible. This limits their usefulness, even when searching for entities, because a person may be referred to in a variety of ways, e.g. by name (Henry VIII), by title (Lord of Ireland), by ancestry (son of Henry VII) or by role (The undersigner). Furthermore, spellings of proper names vary frequently in historical documents (Greiffenclau, Griffenclae), and persons can also be referred to by pronouns. Searching for all variants is not a perfect solution either, since noun phrases such as the King of England can, of course, refer to different entities in different contexts. Finding a particular entity is thus a laborious and error-prone task. If one is interested in certain types of events, for instance all SELLING events [Fertmann, 2013], it becomes even more difficult to locate the desired information. Hence, digitised sources should ideally be enhanced with a semantic-structural annotation layer. Unfortunately, doing this manually is time-consuming, sometimes prohibitively so [Schäfer et al., 2012].

We describe our work of automatically converting a manually compiled collection of historical manuscripts, first into XML-annotated (structured) text and then into a searchable database with a web-based front end. In the conversion process, we exploit typographic and structural cues as well as heuristics based on domain knowledge. The methods we describe are not in themselves entirely novel. Detecting structure in superficially un- or semistructured texts has been the topic of various

research projects. The aim of this paper is thus not so much to introduce novel techniques but to show how relatively simple techniques can be applied to a particular type of collection and thereby make it much more accessible and useful.

2 The Data

The data are a collection of *regesta*, i.e. summaries of medieval charters, pertaining to the City of Saarbrücken and covering a period of 950 years from 601 to 1545 AD [Eder-Stein, 2012]. Compiling this collection, originally on file cards and later electronically in a Microsoft Word document, started in the late 1950s and ended in 2011. Initially, only an open access book publication was intended, hence the use of a Word document to collect the data, which is, of course, suboptimal from a processing point of view.

Since the manuscript was intended for off-line reading rather than electronic processing, the structure of each regest remained largely implicit, being signalled mainly by typographical means. Figure 1 shows an example. The first line is printed in bold face and contains the year and sometimes also the place of issue as well as additional information, e.g. regarding the reliability of the dating. This line is also used as a unique identifier for the regest in the collection. Then follows the modern German summary of the manuscript. Some passages, often names, are left in their original form and printed in italics, e.g. *Sarebrugka debellatur*. Various types of metadata follow, including the original dating, signatories, and archival information. Not all metadata are available for each regest but if they occur, their order is fixed and they are usually preceded by keywords such as *Druck* (print).

1009

König Heinrich II. erobert auf seinem Kriegszug gegen den Bischof [Dietrich II.] von Metz im Spätsommer bis Herbst die Burg Saarbrücken (*Sarebrugka debellatur*).

Annales Altahenses majores

Druck: MGH SS 20 (1868) S. 790; MGH SSRG 4 (1891) S. 16
Regest: MRR I (1876) S. 335 Nr. 1181; Jungk (1914/19) S. 12 Nr. 37;
RIHeinrichII (1971) S. 960 Nr. 1716a - (Ed)

Fig. 1: Regest

The book also contains an extensive index, which not only lists named entities (NEs) referred to in the texts but also provides valuable additional information, e.g. about alternative spellings, family relationships (Metze, Witwe des Schultheiß Nikolaus "Metze, widow of ..."), relationships between locations and person (Einwohner "residents"), titles and roles of persons (Ritter "knight", herrschaftlicher Schneider "stately taylor"), localisation of place names (Kelz, Dorf (Dep. Haute-Saône, F.)), and contexts in which an entity was mentioned in the text (Besuch in Saarbrücken "visit to Saarbrücken"). Figure 2 shows three index entries.

3 Related Work

Several projects are dedicated to digitising medieval charters and making them available via sophisticated web interfaces. The most well-known is *Regesta Imperii*, 1 which provides electronic access to a collection of charters from the period of the Holy Roman Empire [Kuczera, 2005]. Another project is the Charters Encoding Initiative, which has been running since 2004.2 However, as far as we know, in all of these projects, the underlying database is built manually rather than by extracting information from existing texts.

On the other hand, there is a large body of work concerned with determining structure in texts using supervised [Borkar and Sarawagi, 2001, Viola and Narasimhand, 2005] or unsupervised [Grenager et al., 2005] machine learning as well as bootstrapping from existing resources [Canisius and Sporleder, 2007]. Our methods are not as sophisticated nor do they have to be since we can infer a lot of information from typographic cues and domain knowledge. Also related are studies which aim at identifying structure in published papers

[Schäfer et al., 2012]. Typically, these, too, employ heuristics [Schäfer and Weitz, 2012].

```

Schweinheim gen. von Steinbach, Familie von
  Konrad, Edelknecht 1376-01-27, 1374-08-15 (c), 1377-10-20
  Heinrich 1453-08-08
Schwerdorff/Schwerdorff, Dorf (Dep. Moselle, F)
  Güter 1369-07-24
Seewiller (=Seelweiler ?, Wüstung bei Saarlouis; Staerk, Wüstungen, Nr.
  353)
  Einwohner
  - Konrad 1538-08-03
  Haus 1538-08-03

```

Fig. 2: Index Entries

4 From Unstructured Text To Searchable Database

The structure of regest texts and index entries is only implicitly encoded by formatting (type face, level of indentation) and ordering of elements. Many of these devices are ambiguous, e.g. italics are predominantly used to indicate original passages but metadata information is also sometimes set in italics (e.g. vgl. "cf."). We implemented heuristics to determine the function of a piece of text depending on its typographical properties and position relative to other text elements. We also made use of a limited set of manually supplied keywords, such as titles, honorifics and words for groups of persons, mainly for parsing the index entries. This procedure should be relatively easily adaptable for other, similar collections. Once the main structural blocks had been identified, index and regest entries could be linked. Because the index lists NEs and links them to the regesta, we could also relatively easily identify and disambiguate the entities referred to in the regesta themselves, thus avoiding a separate NE recognition step.

```

<location-header>
  <placeName>
    <settlement type="Dorf" abandoned-village="true" av-ref="Staerk, Wüstungen Nr. 19">
      Arschofen</settlement>
    <ddNames> (
      <ddName>Arshofen</ddName>
    </ddNames>, Dorf
    <reference-point>im Köllertal</reference-point> (Wüstung,
    <district>Gde. Gerweiler, Stadtverband Sb.</district>,
    <region type="Bundesland">SL</region>; Staerk, Wüstungen Nr. 19
  </placeName>
</location-header>

```

Fig. 3: XML Mark-Up in the Index

An XML-schema was designed for representing the collection. We tried to comply with the TEI guidelines [Burnard and Bauman, 2013] whenever possible, however, since TEI does not explicitly cover medieval charters we had to deviate from it occasionally. In general, we designed the schema in such a way that it is extensible. In particular, all automatically inferable information types should be encodable, even if we do not extract them immediately. For example, we refrained from automatically identifying the issuer of a document. However, this can normally be identified relatively reliably as it is typically the first person named in the text. Figure 3 shows part of an (instantiated) example of an index entry of type LOCATION. We encode the type of settlement (Dorf "village"), whether it is abandoned or not, the name and alternative names, and the area, district and region, if provided.

Using the heuristics, an XML marked-up version of the summaries and index was automatically generated from the original Word document. Additionally, the extracted information was stored in a database. We consider the XML file the primary format. However, a database is useful because it can be employed as a straightforward backend for a web application. Furthermore, if the collection is extended at a later stage, a database offers a simple interface for data entry. Additions to the collection can be entered directly in the database, thus rendering the automatic conversion from unstructured text to structured XML unnecessary. The updated database can easily be exported to XML and the proofs for future editions of the book can be generated directly from the XML file. Finally,

we created a web interface³ for searching and browsing the collection, which also implements additional features, such as a time line, allowing users to see all regesta from a pre-selected period.

5 Conclusion and Future Work

Where originally only keyword search could be performed on the collection, automatic processing greatly enhanced its utility. Linking of NEs to index entries now allows entity-based search. Furthermore, users can search explicitly for certain information types, e.g. archival information. The explicit mark-up of the documents makes it trivial to implement more complex, logical search options such as searching for co-occurring entities. It is also possible to link to external resources such as maps or show family trees based on the index information. Finally, the database makes the data more easily extensible and allows for error and consistency checking.

With the basic structure in place, one can go further and identify co-reference chains, semantic argument structures or document topics, or extract information, e.g. about all financial transaction in which a given monastery was involved or about the contexts in which certain groups of people are mentioned.

The enhanced collection is not only useful for historians. It is also a valuable resource for students and pupils studying medieval or local history. Historical linguists also have expressed an interest in using it to study historical place names. Moreover, we expect a significant demand from historically inclined laypersons, especially those living in the region covered by the collection.

References

- Kaustubh Deshmukh Borkar and Sunita Sarawagi** (2001). *Automatic segmentation of text into structured records*. In Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data, pages 175–186.
- Lou Burnard and Syd Bauman** (2013), editors. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Text Encoding Initiative Consortium Charlottesville, Virginia.
- Sander Canisius and Caroline Sporleder** (2007). *Bootstrapping information extraction from field books*. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pages 827–836, Prague, Czech Republic, June 2007. Association for Computational Linguistics.
- Irmtraud Eder-Stein, editor** (2012). *Regesten zur Geschichte der Stadt Saarbrücken (bis 1545)*. Publikationen der Saarländischen Universitäts- und Landesbibliothek. universaar: Universitätsverlag des Saarlandes. Bearbeitet unter Verwendung von Vorarbeiten von Hanns Klein. **Susanne Fertmann** (2013). *Extraction of selling events from historical documents*. Bachelor Thesis, Saarland University. Trond Grenager, Dan Klein, and Christopher Manning. Unsupervised learning of field segmentation models for information extraction. In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05), pages 371–378, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.
- Andreas Kuczera** (2005). *Die Regesta Imperii Online*. In Deutsche Kommission für die Bearbeitung der Regesta Imperii e.V. bei der Akademie der Wissenschaften und der Literatur in Verbindung mit der Bayerischen Staatsbibliothek in München, editor, Workshop Buch und Internet - Aufbereitung historischer Quellen im digitalen Zeitalter, pages 3–4.
- Ulrich Schäfer and Benjamin Weitz** (2012). *Combining OCR outputs for logical document structure markup*. Technical background to the ACL 2012 contributed task. In Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries, pages 104–109, Jeju Island, Korea, July 2012. Association for Computational Linguistics.
- Ulrich Schäfer, Jonathon Read, and Stephan Oepen** (2012). *Towards an ACL Anthology Corpus with logical document structure*. An overview of the ACL 2012 contributed task. In Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries, pages 88–97, Jeju Island, Korea, July 2012. Association for Computational Linguistics.
- Paul Viola and Mukund Narasimhand** (2005). *Learning to extract information from semistructured text using a discriminative context free grammar*. In Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 330–337.

Interdisziplinarität modellieren – Über die Modellierung einer Ontologie wissenschaftlicher Prozesse für den Exzellenzcluster Bild Wissen Gestaltung

Stein , Christian

Humboldt-University Berlin, Germany, Cluster of Excellency Image Knowledge Gestaltung

Im November 2012 ist an der Humboldt Universität zu Berlin der Exzellenzcluster Bild Wissen Gestaltung gestartet, der sich selbst als „interdisziplinäres Labor“ bezeichnet. Die Idee hinter diesem Großprojekt ist es, unterschiedlichste wissenschaftliche Disziplinen in ungewöhnlichen und neuen Zusammenstellungen an Fragen unserer Zeit arbeiten zu lassen. Zu einem geisteswissenschaftlichen Schwerpunkt kommen nicht nur Natur- und Technikwissenschaften zusammen, sondern erstmals gleichberechtigt auch Designer und Gestalter. Insgesamt sind über 25 Disziplinen beteiligt. So arbeiten in einzelnen Basisprojekten beispielsweise Architekten, Kunsthistoriker, Germanisten, Informatiker, Chemiker, Kulturwissenschaftler und Interaction Designer zusammen an einer gemeinsamen Fragestellung. Mit zu Hochzeiten über 200 Mitarbeitern ist damit eine Forschungskonfiguration geschaffen, die einzigartig ist und Gelegenheit bietet, das Zusammenarbeiten und die Kommunikationsstrukturen in solchen massiv interdisziplinären Konstellationen zu untersuchen. Darin liegt eines der Hauptinteressen des Clusters: Was ist Interdisziplinarität? Wo funktioniert sie tatsächlich? Und wo nicht? Wann ist sie wirklich hilfreich und zielführend und wann verkommt sie zu einem Schlagwort? Schließlich: Wie kann man das Gelingen der Kommunikation zwischen den Disziplinen sicherstellen bzw. verbessern? Diesen Fragen nähert sich der Cluster mit einer umfangreichen Selbstbeobachtungsstrategie, die in vielen Teilen technisch umgesetzt wird.

Damit steht eine genuin interdisziplinäre Fragestellung im Zentrum dieses Forschungsgroßprojektes und gleichermaßen eine extreme Kommunikationssituation, die sich nur mit digitalen Mitteln adäquat erfassen lässt. Dabei bekommt die Analyse von verwendeter Terminologie und gegenseitigem Verständnis einen wesentlichen Fokus. Selten jedoch konnte sich bisher auf eine spezifische Definition geeinigt werden, die alle Beteiligten vollständig unterstützen. Auch das kann als Terminologiearbeit verstanden werden – allerdings mit einer völlig anderen Zielsetzung als in der klassischen Terminologiearbeit. Dennoch bleibt es das Ziel, auch dieses Arbeiten an Terminologie zu formalisieren, zu ordnen und beschreibbar zu machen.

Die Basisprojekte *Virtuelle und reale Architektur des Wissens und Shaping Knowledge* setzen sich mit den Möglichkeiten einer solchen Formalisierung auseinander. Als besonders hilfreich haben sich dabei die Erkenntnisse aus dem iglos-Projekt (Akronym für „intelligentes Glossar“) der Technischen Universität Braunschweig erwiesen, mit dem der Exzellenzcluster kooperiert. Das iglos-Team hat jahrelang die Kommunikationsprozesse in interdisziplinären Teams untersucht und darauf basierend ein eigenes Terminologiemanagementsystem entwickelt (www.iglos.de). Dieses setzt auf terminologische Ontologien statt der reinen

Sammlung und Listung von Benennungen, Definitionen und ein paar Metadaten. So wird es möglich, die semantischen Zusammenhänge von terminologischen Einträgen als Netzwerk zu modellieren und zu visualisieren, die für ein tatsächliches Begriffsverständen oft unerlässlich sind. Ein solches semantisches Netz ist auch in der Lage, verschiedene Verständnisräume zu modellieren, die sich nicht ohne weiteres vereinheitlichen lassen.

Für das Erforschen dieser Struktur als Kommunikationssituation wurde daher die Entscheidung getroffen, eine Cluster-Ontologie aufzubauen, die auf die Modellierung einer erweiterten Version der scholarly primitives setzt (Unsworth 2000, Bamboo 2010, Blanke et. al. 2011). Dabei werden Sprachkonstrukte aus bestehenden Ontologien wie beispielsweise Dublin Core, SKOS, FOAF und CIDOC CRM wiederverwendet (Tzompanaki et. al. 2012). Wichtige Überlegung dabei ist, dass sich die Adaption und Verwendungsweise von Terminologie nur dann erfassen lässt, wenn man sich ihre Anwendung durch Personen oder in Quellen und ihre Kookkurrenzen in diesen Umgebungen ansieht. Die Beschreibung von Verständnisräumen erfordert es, die reale Anwendung von Terminologie zu untersuchen und nicht nur die gewollte: Wer verwendet sie? Wer kommuniziert mit wem damit? Welche Quellen verwenden sie? Welche Person rezipiert welche Quellen? Wie sind diese Quellen verbunden? Welche Themen und Communities sind durch welche Terminologie charakterisierbar? Und schließlich: Wer versteht wen unter welchen Umständen überhaupt richtig? Wo kann ein Forscher für ihn interessante Quellen finden? Und mit wem kann er sich über bestimmte Themen unterhalten? Schließlich: Welche basalen Tätigkeiten machen den Alltag eines Forschers aus und welche Konfigurationen davon sind in interdisziplinären Konstellationen überhaupt kommensurabel bzw. kompatibel?

Vier zentrale Top-Level-Entitäten organisieren gemäß den oben genannten Überlegungen das Schema der Ontologie: *Personen*, *Quellen*, *Themen* und *Termini*. Unter Personen sind zunächst die konkreten Mitarbeiter des Clusters zu verstehen; hier werden aber auch angegliederte Personen, virtuelle Akteure sowie Personengruppen, Projekte und Teams modelliert. So entsteht zunächst ein Personennetzwerk das Informationen darüber enthält, wer in welchen Organisationseinheiten mit wem steht. Die kontinuierlich durchgeführte Selbstbeobachtung der Clustermitarbeiter erlaubt aber auch eine Modellierung der Kommunikationsverbindungen zwischen den Personen. So kann angegeben werden, wer wie umfangreich mit wem kommuniziert und auf welchem Wege. Diese Informationen werden teils automatisch mit technischen Mitteln und teils durch empirische Beobachtung erfasst.

Aber nicht nur Personen und deren Verbindungen untereinander werden modelliert, sondern auch die Objekte, mit denen sie umgehen. Diese sind in der Ontologie als *Quellen* gefasst. Quellen können beispielsweise Bücher oder Zeitschriften sein, die Personen lesen. Genauso kann es sich aber auch um produzierte Texte handeln. Neben Texten werden als Quellen auch Zeichnungen, Fotos und Bilder, Modelle und andere Datenformate gefasst. Eine Quelle ist also ein beliebiges, physisches oder digitales Objekt, mit dem interagiert wird. Quellen können auch Beziehungen untereinander aufweisen, beispielsweise im Sinne von Zitation oder Beeinflussung. Auch diese Beziehungen können in der Ontologie detailliert modelliert werden.

Der Sprung auf die inhaltliche Ebene geschieht mithilfe der *Themen*. Themen fungieren als zusammenfassende Charakterisierung von zusammenhängenden Inhalten. Typische Themen sind beispielsweise Sachgebiete, Domänen, Ausbildungsgänge und ähnliches. Themen können in Baumstrukturen aufgebaut werden, so dass sich allgemeinere und speziellere Themen anlegen und miteinander verbinden lassen. So kann man Themen zur Charakterisierung von Kompetenzen von Personen verwenden. Oder sie werden genutzt, um Quellen zu klassifizieren und so Aussagen über die darin behandelten Inhalte zu treffen. Dabei ist es wahrscheinlich, dass Personen oder Quellen jeweils mehrere Themen zugeordnet bekommen. Die Flexibilität der Ontologie ermöglicht es, verschiedene existierende Klassifizierungssysteme zu integrieren. Hier wird beispielsweise

auf das Open Directory zurückgegriffen werden, das eine der umfangreichsten Themenklassifikationen beinhaltet.

Der aus terminologischer Perspektive vielleicht interessanteste Teil der Ontologie ist der des *Terminus*. Unter *Terminus* wird im Folgenden „das zusammengehörige Paar aus einem Begriff und seiner Benennung als Elemente einer Terminologie“ verstanden. Da Benennungen sehr häufig in unterschiedlichen Varietäten vorkommen, dort jeweils jedoch mit unterschiedlichen Begriffen verbunden werden, erlaubt die anvisierte Modellierung eine Strukturierung auch dieser Fälle.

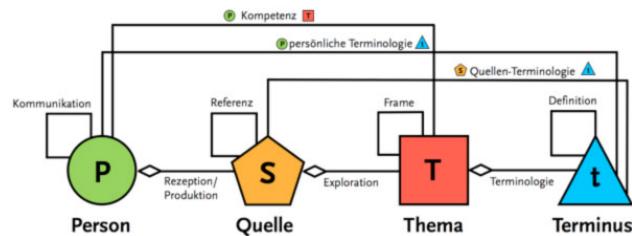


Fig. 1:

In der Ontologie des Clusters werden Termini auf unterschiedlichen Ebenen modelliert. Zum einen wird den einzelnen Basisprojekten die Möglichkeit gegeben, terminologische Festlegungen für ihre Bereiche vorzunehmen und sich über die anderer zu informieren. Zum anderen werden in größeren Arbeitsgruppen die oben genannten „großen Begriffe“ diskutiert und theoretisiert werden. Die in diesem Rahmen zu diskutierenden Benennungen beginnen mit dem Namen des Clusters selbst – *Bild*, *Wissen*, *Gestaltung* und *Interdisziplinarität*. Andere wichtige große Begriffe formieren sich um die Benennungen *Struktur*, *Modell*, *Architektur*, *RaumoderCode*. Es leuchtet schnell ein, dass diese Bemühungen zwar auch kurze und prägnante Definitionen anstreben, diese aber zwingend umfangreicher Erläuterungen bedürfen. Darüber hinaus reicht es in diesen Fällen nicht aus, eine einzelne Definition zu finden – vielmehr ist es erforderlich eine ganze Menge dazugehöriger Termini zusammenzubringen, zu strukturieren und zu definieren. Für ein solches Vorgehen sind Ontologien besonders geeignet, da sie die Verbindungen und Zusammenhänge zwischen den Termini modellieren können. Der dritte Anwendungsbereich erstreckt sich auf die Analyse tatsächlicher Terminologieverwendung in den Quellen des Clusters. Ein Großteil der verwendeten Textquellen wird wenn nötig digitalisiert und zentral verwaltet werden. Dazu kommt das Open-Source-System Zotero zum Einsatz, das eine direkte RDF-Schnittstelle besitzt. So kann aus der Ontologie direkt auf die Quellen verlinkt werden. Mit verschiedenen Textminingmethoden werden die Digitalisate dann untersucht. Dazu gehört beispielsweise die Topic Detection, die es ermöglicht, Quellen bestimmten Themen zuzuordnen und den Grad der Relevanz der Quelle für ein Thema zu berechnen. Den Themen in der Ontologie sind dazu Termini zugeordnet, die signifikant für das jeweilige Thema sind. Die Ontologie lernt über die Analysen automatisch neue Termini, die zu einem gegebenen Thema gehören.

Alle diese Entitäten und die Relationen zwischen ihnen werden je nach Bedarf in Unterklassen spezialisiert.

Das Ziel ist es dabei nicht, individuelle und disziplinspezifische Arbeitsweisen über einen Kamm zu scheren – im Gegenteil: Gerade die Heterogenität der Arbeitsweisen besser zu verstehen und mit einer empirischen Datenbasis zu versehen ist nach unserer Überzeugung die Voraussetzung für eine bessere, vielfältigere und innovativere Forschungsarbeit. Wenn sich die Rahmenbedingungen interdisziplinärer Forschung auf Basis der gewonnenen Erkenntnisse auch nur etwas verbessern lassen, ist bereits viel gewonnen – denn interdisziplinäre Strukturen werden immer häufiger und nichts verschwendet so viele Ressourcen wie ungünstige Bedingungen und falsche Annahmen.

Am 1. Juni fand die offizielle Eröffnung des Exzellenzclusters *Bild Wissen Gestaltung* statt. Hunderte Besucher haben die Projekte begutachtet, sich informiert und angeregt diskutiert. Nun heißt es für unser Team, unsere Pläne umzusetzen. Was daraus wird und wie sich unsere Arbeit entwickelt kann man

mit kontinuierlichen Updates auf der Webseite des Clusters verfolgen unter www.interdisciplinary-laboratory.hu-berlin.de.

References

- Bamboo** (2010): Project Bamboo Scholarly Practice Report. wikihub.berkeley.edu/download/attachments/68619618/bamboo_scholarly_practice_report.pdf?version=2&modificationDate=1292887646732
- Blanké, T., & Hedges, M.** (2011). *Scholarly primitives: Building institutional infrastructure for humanities e-Science*. Future Generation Computer Systems. doi:10.1016/j.future.2011.06.006
- Tzompanaki, K., & Doerr, M.** (2012). *Fundamental Categories and Relationships for intuitive querying CIDOC CRM based repositories*. 139.91.151.170/tech-reports/2012/2012.TR429_Intuitive_querying_CIDOC-CRM.pdf
- Unsworth, J.** (2000). *Scholarly Primitives: what methods do humanities researchers have in common, and how might our tools reflect this?* Symposium on Humanities Computing formal methods experimental practice. www3.isrl.illinois.edu/~unsworth/Kings.5-00/primitives.html

Cirilo Client: An application for data curation and content preservation

Steiner, Elisabeth
elisabeth.steiner@uni-graz.at
 University of Graz

GAMS: A Fedora Commons instance

Since 2003 the Centre for Information Modeling - Austrian Centre for Digital Humanities at the University of Graz (Austria) provides an infrastructure for a variety of DH projects. After years of building insular solutions, the Centre introduced a powerful yet flexible new infrastructure, called GAMS (Geisteswissenschaftliches Asset Management System, AMS for the Humanities). It is based on the Fedora Commons architecture. Thus, the infrastructure inherits all features already provided by Fedora: full OAIS-compliance, strict separation of data and metadata, and predefined interfaces like OAI-PMH. A central advantage of the Fedora architecture is its object model: An asset consists of a primary source, some metadata and virtual representations derived from the primary source. The object is completely self-descriptive: It knows about all changes that have been made to it, its version history, datastreams and assigned context objects. Finally, it also knows about all possible representation forms. Each object contains all the necessary information to store, preserve, retrieve and view it.

Cirilo Client: Mass operations in Fedora made easy

Although Fedora is a powerful tool, front-end object management is not always easy, especially with regard to mass operations. The Centre has developed a tool for this use case, complementing Fedora's built-in Admin Client. Cirilo is a java application developed for data curation and content preservation in Fedora-based repository systems. Content preservation and data curation in our sense include object management and creation, versioning, normalization and standards, and choice of data formats.

Cirilo makes use of Fedora's management-API (API-M). It offers applications which are particularly prone to being used as tools for mass operations on Fedora repository objects, such as ingest or replacement processes: With Cirilo ingest processes can be performed from the file system, from an eXist database or an Excel spreadsheet. During the ingest metadata is automatically extracted from the source document and written to the newly created object (for instance in DC format).

The client operates on a collection of predefined content models which can be used without further adjustments for standard workflow scenarios like the management of collections of TEI objects. The content models, which are based on the Fedora object model, are class definitions: On the one hand they define the (MIME-)type of the contained data streams, on the other hand they designate dissemination methods operating on these data streams. Every object in the repository is an instance of one of these class definitions. The advantage of this concept lies in the fact that very complex data sources and workflows can be handled easily.

Currently, the client offers various content models for specific purposes, special emphasis lies on the TEI model. The TEI ingest processes can be flexibly customized: during ingest policies for the extraction of semantic information can be applied, referenced images can be uploaded simultaneously and ontology concepts can be resolved. A new content model currently in development creates the appropriate ontology objects, especially SKOS objects. A designated query object makes it possible to pose queries with parameters to the Mulgara triplestore. With the help of these ontology and query objects dynamic indices can be created. There is a container object for the creation of collections available, which makes it easy to organize your resources. Finally, there are some models optimized for specified primary sources like METS/MODS, HTML, PDF, BibTeX or external resources accessible via an URL. A content model for linguistic resources is in development (in cooperation with ICLTT, Vienna). Currently, we are testing how controlled vocabularies and thesauri (for instance geonames.org), can be sensibly integrated in the system.

The user can assign numerous virtual representations via the client. The METS/MODS object is designed to be viewed in the DFG-Viewer. TEI objects can be directly used as the input for the Voyant Tools or the Versioning Machine. The members of a context object can be projected on a map using Google Maps. Basically any web-based service can be integrated into the infrastructure. Of course, user- and project-specific stylesheets are often employed.

The Cirilo Client will be made available as an open source software project, including documentation, as a contribution of the Centre for Information Modeling - Austrian Centre for Digital Humanities to Dariah-AT in 2014.

References

- Dariah-EU: www.dariah.eu [2013-10-28]
 DFG-Viewer: dfg-viewer.de/ueber-das-projekt [2013-10-28]
 Fedora Commons: www.fedora-commons.org [2013-10-28]
 Google Maps: maps.google.at [2013-10-28]
 Geisteswissenschaftliches Asset Management System, AMS for the Humanities: gams.uni-graz.at [2013-10-28]
Carl Lagoze, Sandy Payette, Edwin Shin, Chris Wilper, Fedora (2005). An Architecture for Complex Objects and their Relationships. arxiv.org/ftp/cs/papers/0501/0501012.pdf [2013-10-28]
 Versioning Machine: v-machine.org [2013-10-28]
 Voyant Tools: voyant-tools.org [2013-10-28]

The DigiPal Framework for Script and Image

Stokes, Peter A.
peter.stokes@kcl.ac.uk
 King's College London

Brookes, Stewart
stewart.brookes@kcl.ac.uk
 King's College London

Noël, Geoffroy
geoffroy.noel@kcl.ac.uk
 King's College London

Buomprisco, Giancarlo

giancarlo.buomprisco@kcl.ac.uk

King's College London

Marques de Matos, Debora

debora.matos@kcl.ac.uk

King's College London

Watson, Matilda

matilda.watson@kcl.ac.uk

King's College London

An on-going challenge in palaeography is, as Albert Derolez expressed it, how arguments about ancient and medieval handwriting can be 'as clear and convincing to their reader as to their author'.¹ Derolez suggested greater use of quantitative evidence, and his suggestion has since been embraced by groups working on automated or semi-automated identification of scribal hands or styles of handwriting.² Such methods are almost entirely statistical, using image processing techniques to extract metrics and to apply those in supervised or unsupervised ways. However, these methods have very rarely been used by palaeographers who are naturally oriented towards symbolic or semantic (i.e. verbal) approaches. Indeed some palaeographers have rejected quantitative methods on principle, such as Armando Petrucci who has argued that they 'cannot simply exist' (*non può semplicemente esistere*) in a fluid and human context like handwriting.^{3 4} More sympathetic scholars have still raised questions about trust in what they see as 'black boxes', asking how to validate these more automated methods, how to test the assumptions underlying them, and how these approaches can be convincing to an audience which cannot reasonably be expected to understand them.^{5 6} Some few have therefore taken an alternative, more symbolic or semantic approach, looking to the computer not to provide 'answers' but instead to allow manipulation and exploration of material in a more intelligible, if less conclusive, manner.^{7 8 9} This reflects a wider argument that the computer should not be used to provide answers for Humanities researchers, but rather should suggest, stop short, and allow people to construct knowledge and understanding through active manipulation and visualisation.^{10 11} However, such symbolic approaches are rare and have also gained no acceptance in the past.

This challenge to provide 'clear and convincing' palaeographical descriptions has been taken up by DigiPal, a four-year project funded by the European Research Council (FP7 Grant Agreement No. 263751). One early outcome has been a formal model for describing handwriting, the first of its kind and which is already changing palaeographical descriptions.¹² Since then, the model has been generalised and implemented in an open-source web-based framework. Instead of relying on image processing or automated methods, team members manually annotate images with highly structured descriptions.¹³ The framework is being used in DigiPal for eleventh-century English Vernacular minuscule, but also for 'ScandiPal' on twelfth-century Latin from Norway and Sweden, 'SephardiPal' on fifteenth century Hebrew from the Iberian Peninsula, RIM on early medieval coins, and 'Models of Authority' on twelfth-century Scottish charters. It has also been extended to decoration, thereby helping Art Historians who have faced similar difficulties in their own descriptions and arguments. The existing DigiPal site currently presents records of all the (approximately) 1,200 known surviving examples of eleventh-century writing in Old English, along with images and structural annotations of more than half of them. This is useful not only for palaeographical research but also for teaching, as students or the interested public can (for example) look at a page of handwriting from the area where they live in England, or highlight letters on the page to help them learn to read the documents. Examples from the site are shown below.

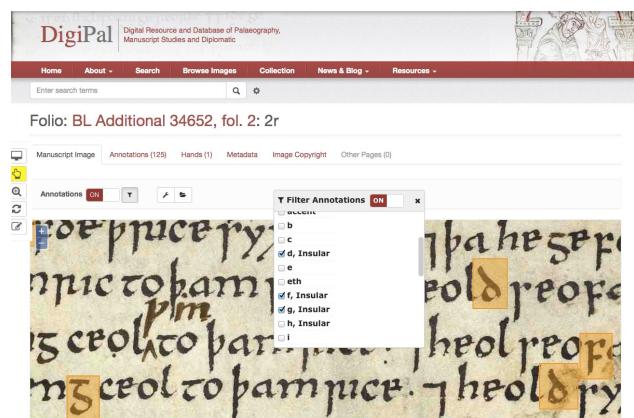


Fig. 1: Annotations of Insular d, f and g in an example image of a manuscript page. Screenshot of <http://digipal.eu/digipal/page/362/>

A screenshot of the DigiPal search interface. The search bar at the top contains the query 'Eadwig Basan'. Below the search bar are several filter dropdowns: 'Script' (set to 'Manuscripts'), 'Character' (set to 'Allograph'), 'Allograph' (set to 'descender'), and 'Feature' (set to 'Feature'). The results section shows a grid of small thumbnail images of manuscripts. Each thumbnail includes basic metadata: 'Confirmation of Privileges (44v)', 'Hand' (Main Hand), 'Scribe' (Eadwig Basan, saec. xi/4), 'Place' (CaCC Secc. xi/4), 'Date' (British Library), 'Repository' (Royal), 'Shelfmark' (G. 447 K. 247), and 'Catalogue Number' (38). The interface also includes tabs for 'Images' and 'List'.

Fig. 2: Examples of letters with descenders written by 'Eadwig Basan', a scribe of Christ Church Cathedral, Canterbury. Screenshot of http://digipal.eu/digipal/search/?terms=Eadwig+Basan&view=list&basic_search_type=graphs&component=descender

A screenshot of the DigiPal interface for describing letterforms. The interface is divided into two main sections: 'd, Insular' on the left and 'f, Insular' on the right. Each section shows a grid of small thumbnail images of the letter forms. To the right of each grid is a detailed list of descriptive checkboxes. For 'd, Insular', the list includes: 'concave left', 'concave right', 'c-shaped', 'diamond-shaped', 'flat-topped', 'homed', 'narrow', 'open', 'short', 'stem', 'backward-reaching', 'bilinear', and 'bifurcating'. For 'f, Insular', the list includes: 'concave left', 'concave right', 'c-shaped', 'diamond-shaped', 'flat-topped', 'homed', 'narrow', 'open', 'short', 'stem', '45°', 'back vertical', 'bifurcating', 'broken', 'curved down', 'concave down', 'concave up', 'long', 'palm-topped', 'round', 'semi-circular', 'square', 'teardrop-shaped', 'vertical left', 'straight', 'tracking left', 'tracking right', 'turned down', 'turned right', 'vertical', 'vertical lip', and 'wedged'. A 'Saved annotation' button is located at the bottom of the interface.

Fig. 3: Interface for describing letterforms (available to team members only). Screenshot of <http://digipal.eu/digipal/page/80/allographs/>

The project has already received substantial attention. In the eleven months since its launch, the new prototype has received over 15,000 visits, with over 45,000 page views and 9,000 'unique' visitors; it has been referred to in at least eleven blogs, including those by the British Library¹⁴ and the Bishop of Huntingdon¹⁵; it has been cited in at least five scholarly articles (excluding those by the project team)^{16 17 18 19 20}; and it is being used for teaching in the UK and abroad. This strongly suggests that the project has achieved an important goal in being accessible to palaeographers, medievalists and the public in a way that other approaches have not.

The proposed poster will present the model for describing handwriting and decoration, as well as further details about the framework, its models and implementations. A live demonstration will be available not only of the DigiPal system but also ScandiPal and SephardiPal, two projects which are not publicly visible because of limitations in image rights. The

framework is available already for download from GitHub²¹, and we are very happy to discuss its customisation for other projects.

References

1. **Derolez, A.** (2003). *The Palaeography of Gothic Manuscript Books* (Cambridge: Cambridge University Press), pp. 7–8.
2. **Hassner, T. et al.** (2013). *Computation and Palaeography: Potentials and Limits*. Dagstuhl Manifestos 2: 14–35. doi:10.4230/DagMan.2.1.14
3. **Costamagna et al.** (1995 and 1996). *Commentare Bischoff*. Scrittura e Civiltà 19: 325–48 and 20: 401–7.
4. **Pratesi, A.** (1998). *Commentare Bischoff: un secondo intervento*. Scrittura e Civiltà 22: 405–8.
5. **Stokes, P.A.** (2012). *Palaeography and the 'Virtual Library'*. In Nelson, B., and Terras, M. (eds). *Digitizing Medieval and Early Modern Material Culture*. Arizona: Center for Medieval and Renaissance Studies. 137–69.
6. **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. doi:10.1093/lit/fqn019
7. **The Cuneiform Digital Palaeography Project** (2009). www.cdp.bham.ac.uk/ (last accessed 3 March 2014).
8. **Stokes, P. A., et al.** (2011–14). *DigiPal: Digital Resource and Database of Palaeography, Manuscripts and Diplomatic*. London: King's College. digipal.eu/ (last accessed 3 March 2014).
9. **ManCASS** (2005). *C11 Database of Scripts and Spelling*. Manchester: University of Manchester. web.archive.org/web/20090520145856/http://www.arts.manchester.ac.uk/mancass/c11database/ (last accessed 3 March 2014).
10. **Chang, D. et al.** (2009). *Visualizing the Republic of Letters*. Stanford: Stanford University. www.stanford.edu/group/toolingup/rplviz/papers/Vis_Rofl_2009 (last accessed 3 March 2014).
11. **Clement, T. et al.** (2009). *How Not to Read a Million Books*. Waltham, MA: Brandeis University. people.brandeis.edu/~unsworth/hownot2read.html (last accessed 3 March 2014).
12. **Stokes, P.A.** (2012). *Modelling Medieval Handwriting: A New Approach to Digital Palaeography*. In DH2012 Book of Abstracts. Ed. J.C. Meister et al. Hamburg: University of Hamburg. 382–85. www.dh2012.uni-hamburg.de/conference/programme/abstracts/modeling-medieval-handwriting-a-new-approach-to-digital-palaeography (last accessed 3 March 2014).
13. **Stokes, P.A.** (2013). *'What, No Automation?' Some Principles of the DigiPal Project*. In DigiPal: Digital Resource and Database of Palaeography, Manuscripts and Diplomatic. London: King's College. digipal.eu/blog/what-no-automation-some-principles-of-the-digipal-project/ (last accessed 4 March 2014).
14. **Harrison, J.** (2013). *Have you used DigiPal Yet?* Medieval Manuscripts Blog. London: British Library. britishlibrary.typepad.co.uk/digitisedmanuscripts/2013/05/have-you-used-digipal-yet.html (last accessed 3 March 2014).
15. **Thompson, David** (2013). *The Return of the Palaeographers!!* Bishop's Blog. bpdt.wordpress.com/2013/06/07/the-return-of-the-palaeographers/ (last accessed 3 March 2014).
16. **Lowe, K.** (2012). *From Quill to T-PEN: Palaeography, Editing and their E-Futures*. Literature Compass 9:12. 1004–1009. doi:10.1111/lic3.12014
17. **Lee, S.D.** (2012). *Anglo-Saxon Studies and Digital Technologies: Past, Present and Future*. Literature Compass 9:12. 996–1003. doi:10.1111/lic3.12015
18. **Varila, M.** (2013). *Graphetic Variation within One Scribal Hand as Evidence of Manuscript Production*. Studia Neophilologica (Advanced Access). doi:10.1080/00393274.2013.834107
19. **Faulkner, M.** (2012). *Rewriting English Literary History 1042–1215*. Literature Compass 9:4. 275–91. doi:10.1111/j.1741-4113.2011.00867.x
20. **McCarty, W.** (2012). *The PhD in Digital Humanities*. In B.D. Hirsch, ed., *Digital Humanities Pedagogy: Practices, Principles and Politics*. Cambridge: Open Book. 33–46. www.openbookpublishers.com/reader/161 (last accessed 4 March 2014).
21. **Noël, G., et al.** (2012–14). *Digital Resource and Database of Paleography, Manuscripts and Diplomatic*. GitHub. github.com/kcl-ddh/digipal (last accessed 3 March 2014).

Digital multi-text editions from scratch to electronic performance. Transcription and collation routines transformed in a flexible database system

Stolz, Michael

michael.stolz@germ.unibe.ch

University of Bern, Switzerland

The production of digital multi-text editions requires a wide range of preparatory steps. In the case of texts transmitted in (medieval) manuscripts, the witnesses have to be transcribed according to specific encoding rules. The transcriptions then are collated following certain ideas and concepts of how the transmission process could have developed (phylogenetic analysis can help in this concern, cf. Howe et al. 2004 and 2012, Stolz 2013). The transcriptions and collations finally have to be transferred to a digital edition that allows the users to explore the characteristics of single witnesses as well as the history of a text, which is delivered in variants and in different versions. A dynamically organized database offering various components and adapted to the needs of diverse user-profiles is nowadays the right tool for this purpose.

The poster demonstrates the steps described above that are abstracted from the experiences made in the Swiss *Parzival Project*, based at the university of Bern (cf. http://www.parzival.unibe.ch). The electronic edition of Wolfram von Eschenbach's German Grail novel, written shortly after 1200 and transmitted during several centuries in ca. hundred witnesses (complete manuscripts as well as fragments), has now been completed by more than a half of the textual corpus (cf. Stolz 2003 and 2011, Viehauser 2008). During the last years, transcription rules have been established that consider particular manuscript features as well as the compatibility with international standards such as TEI (cf. Stolz et al. 2007). Following these rules, the manuscript transcriptions are made.

The next step after transcription is the collation of the manuscript texts. In the editing process of Wolfram's *Parzival* this decisive stage has to cope with the existence of different textual versions ('Fassungen') created in the author's context (caused by the oral delivery of texts in the Middle Ages). Due to this fact not only different manuscript texts (on a first level), but also the textual versions (founded on different manuscripts groups, on a second level) have to be collated to each other. The poster shows the problems resulting from this two-level collation (performed in the *Parzival* project so far by electronic, but not automatized procedures), and discloses options for resolving them in a future semi-automatic process of electronic text comparison.

At an early stage of the project, program packages such as *Collate* (cf. Robinson 1994) seemed to be an adequate tool for producing an electronic edition right from the transcription up to the final product including collations, variant apparatuses and digital images. In a later period of the project the editors switched to modules supported by components of *TuStep* (Tübinger System von Textverarbeitungs- programm), which, due to its flexible character, fitted in better with their requests. An electronic mask created with *TuStep* allows the editors to handle the complexity featured in the *Parzival* text. However, by using this tool, the collations concentrate more on the editor's handwork than they would have to with the

semi-automatic, yet imperfect techniques offered by *Collate*. Nowadays, program packages such as CollateX (cf. <http://collatex.net/>, accessed 05/03/14) offer alternative tools, but still don't fully satisfy the project's needs. The examples given in the poster show some basic requirements occurring in the collation of a text transmitted in multiple versions. They result in a plea for enhancing the ingenious potential contained in currently obsolescent programs such as *Collate* and *TuStep*.

The poster also demonstrates the concept of the project database that is currently under construction. This tool will enable the users to find their way through a complex textual tradition. They can browse through both single manuscripts and different versions, and they can edit the text in different encodings, both in electronic and printed form.

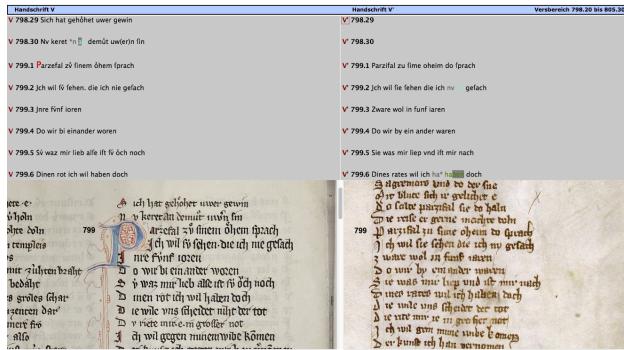
The readers can, in fact, access synoptic listings of single verses in the variant wordings of the different manuscripts, e.g. verse 249.27:

Verssynopse zu Vers 249.27

Vorheriger Vers: 249.26

Vers	Handschrift	Versinhalt
249.27	D	frowe mir ift vil leit.
249.27	m	Frouwe mir ift fere leit
249.27	n	Ffrouwe mir ift fere leit
249.27	o	Frouwe mir ift fere leit
249.27	G	nv wizet frowe mir ift leit.
249.27	I	Nu wizzet frowe mir ift leit.
249.27	L	Nv * wifcent frowe mir ift leit
249.27	M	Nwifzett vrouwe myr ift leit
249.27	O	Nwizzet Frowe mir ift leit.
249.27	Q	Nü wiffet fraw mir ift leyt
249.27	R	Nun wiffent frowe mir ift leid
249.27	T	er fp[er]ach r[ec]hwe mir ift leit
249.27	U	Er fagete* fagete[re] v[er]ewie mir ift leit
249.27	V	Er fagete vrowe mir ift leit
249.27	W	Vil felig frowe mir ift lait
249.27	Z	Nwizzet Frowe mir ift leit
249.27	Fr21	Nv v[er]et frō mir ift leit.-
249.27	Fr36	:;w wiz: fraw: mi::
249.27	Fr40	nv wizzet frowe :;w mir ift leit .
249.27	Fr51	Vrowe mir ift :;de leyt
249.27	Fr69	Frowe mir ift fere leit

They can compare the transcriptions and digitized images of singular manuscripts, e.g., V and V' representing an exemplar and its copy (with V' omitting a substantial passage before 799.1):



And the readers can explore the textual varieties of a critical electronic edition presenting the *Parzival* text in four versions. Links to the relevant manuscripts on which the versions are based on are provided (with an example of ms. D in the second image).

Handschriften: D H N R G J L M O Q R T U V W Z F121 F126 F140 F151 F169	Download: Einstellungen [222] 249 [2010] Startseite
249.1 Der valchite widersaz	* 249.1 der valchite widersaz
249.2 kerte iſ der hofliege kraz.	* 249.2 kerte iſ der hofliege kraz.
249.3 sin schieden, das rnew mich.	* 249.3 sin schieden, das rnew mich.
249.4 aleſt n̄ dventurēt ez sich.	* 249.4 aleſt n̄ dventurēt ez sich.
249.5 dō begunde krenken sich iſ spor.	* 249.5 dō begunde krenken sich iſ spor.
249.6 sich schieden, die dā ritn vor.	* 249.6 sich schieden, die dā ritn vor.
249.7 s̄l war smal, diu ī was breit.	* 249.7 s̄l war smal, diu ī was breit.
249.8 er verſt si gar, das was im leit.	* 249.8 er verſt si gar, das was im leit.
249.9 mare vreich ob de junge man,	* 249.9 Nu vreich der junge man,
249.10 dō von er herzeſt gewan.	* 249.10 mare, dō von er herzeſt gewan.
249.11 dō erholte der degen elens riche.	* 249.11 er verman der heil elens riche
249.12 einer vrouwen stinne jāmerlich.	* 249.12 einer vrouwen stinne jāmerlich.
249.13 ez was danoch von trouwe naz.	* 249.13 er was danoch von trouwe naz.
249.14 vor im ſi einer linden sz.	* 249.14 vor im ſi einer linden sz.
249.15 ein maget, der vuoge iſ trouwe n̄t.	* 249.15 ein maget, der vuoge iſ trouwe n̄t.

Handschrift D: [64] S. 73 [74]
Seite zurück: [66]
Spalte b

249.1 Der valchite widr fazz.
249.2 cherte ſi der hofliege chraz.
249.3 fin fieden und das rnew mich.
249.4 aleſt n̄ Aventurēt fich.
249.5 ob begunde chrenken fich iſ spor.
249.6 fieden di da rnew vor.
249.7 iſ der final dī vaf walt breit.
249.8 er verſt foz dar gaf waf im leit.
249.9 mare vreich do der lvinge man.
249.10 ob von er herzeſt gewan.
249.11 Do erholte der degen elenrich.
249.12 einer vrouwen stinne jāmerlich.
249.13 ez waf danoch von trouwe naz.
249.14 vor im ſi einer linden foz.
249.15 ein maget, der vuoge iſ trouwe n̄t.

The poster concludes with perspectives on the current endeavour of developing sustainable formats for the *Parzival* database and other textual databases in Switzerland, which are performed by a recently established national "Datene- und Dienstleistungszentrum für geisteswissenschaftliche Forschungsdaten" (DDZ), funded by the Swiss Academy of Humanities and Social Sciences.

References

Howe, Christopher J./ Barbrook, Adrian C. / Mooney, Linne R. / Robinson, Peter (2004): *Parallels between Stemmatology and Phylogenetics*, in: Studies in Stemmatology II, ed. by Pieter van Reenen/ August den Hollander/ Margot van Mulken, Amsterdam/ Philadelphia 2004, pp. 3–11

Howe, Christopher J./ Conolly, Ruth / Windram, Heather F (2012): *Responding to Criticisms of Phylogenetic Methods in Stemmatology*, in: Studies in English Literature 1500–1900 52, pp. 51–67

Robinson, Peter: *Collate* (1991). *A Program for Interactive Collation of Large Textual Traditions*, in: Research in Humanities Computing 3. Selected Papers from the ALLC/ACH Conference, Tempe (Arizona), March 1991, ed. by Don Ross/ Dan Brink, Oxford 1994, pp. 32–45

Stoltz, Michael (2003): *New Philology and New Phylogeny. Aspects of a Critical Electronic Edition of Wolfram's Parzival*, in: Literary and Linguistic Computing 18,2, pp. 139–150

Stoltz, Michael/ Schöller, Robert/ Viehhauser, Gabriel (2005): *Transkriptionsrichtlinien des Parzival-Projekts*, in: Edition und Sprachgeschichte. Baseler Fachtagung 2005, ed. by Michael Stoltz/ Robert Schöller/ Gabriel Viehhauser, Tübingen 2007 (Beihefte zu editio 26), pp. 295–328

Stoltz, Michael (2011): *Benutzerführung in digitalen Editionen. Erfahrungen aus dem Parzival-Projekt*, in: Digitale Edition und Forschungsbibliothek. Beiträge der Fachtagung im Philosophicum der Universität Mainz am 13. und 14. Januar 2011, ed. by Christiane Fritze et al., Wiesbaden 2011 (Bibliothek und Wissenschaft 44), pp. 49–80

Stoltz, Michael (2013): *Early versions in medieval textual traditions. Wolfram's Parzival as a test case*, in: Dating Egyptian Literary Texts, vol. 1, ed. by Gerald Moers et al., Hamburg (Lingua Aegyptia – Studia monographica 11), pp. 561–587

Viehhauser, Gabriel (2008): *On the Margins of the Canon – Editions, the 'Whole' Text and the 'Whole' Codex*, in: Variants 7 (2008 [recte 2010]), pp. 57–74

Medical Humanities. Projet de musée digital

Suciuc, Radu
radu.suciuc@unifr.ch
Université de Fribourg

Wenger, Alexandre
alexandre.wenger@unifr.ch
Université de Fribourg

Bolli, Laurent

bolli@bread-and-butter.ch
Agence Bread and Butter

La formation médicale en Suisse doit offrir aux étudiants une plateforme innovante leur permettant de réfléchir aux enjeux sociaux et éthiques du monde de la santé, à l'heure où les nouvelles technologies et le web 2.0 transforment en profondeur l'exercice de la profession médicale. Notre projet consiste en la création d'un site web *responsive*, prenant la forme d'un musée digital accessible via une *webapp* adaptée pour les tablettes et les *smartphones*. Lors du colloque DH 2014 nous présenterons la première exposition virtuelle du musée consacrée aux rapport historiques entre la médecine et l'alimentation. Le poster sera accompagné d'iPads sur lesquels les participants pourront tester la *webapp*. Le musée s'inscrit dans la lignée de projets DH récents qui exploitent des collections muséographiques et favorisent l'interaction des visiteurs à travers des interfaces web (Tales of Things 2009-2013; QRator 2011-2013; Ross 2012).

Ce projet présente pour originalité conceptuelle de se situer à l'intersection de deux champs de savoirs en pleine expansion: d'une part les *Digital Humanities* et, d'autre part, les *Medical Humanities* qui proposent une réflexion interdisciplinaire autour des enjeux sociaux et culturels de la médecine contemporaine (Bates 2013). Par exemple, différents médecins considéraient la pomme de terre comme un aliment révolutionnaire au 18e siècle, capable d'enrayer les famines. Aujourd'hui, l'Organisation des Nations Unies pour l'alimentation et l'agriculture voit dans les insectes un moyen de pallier au manque de nourriture lié à l'augmentation de la population mondiale et à l'épuisement des ressources naturelles (Durst 2012). Les DH apportent de nouveaux modes de circulation entre ces contenus. Les aliments du passé et ceux d'aujourd'hui seront présentés sous une forme ludique et accrocheuse, capable d'intéresser un public large à des contenus de niveau académique. L'apport des méthodes et des outils DH s'avère ainsi particulièrement adapté et fécond pour organiser, enrichir et diffuser le type de savoirs véhiculé par les *Medical Humanities*.

Du point de vue des outils informatiques utilisés, le projet profite des dernières avancées disponibles dans les technologies web (HTML5, PHP 5 Orienté objet, JavaScript Orienté objet, jQuery, S-CSS). La mise en place d'un Content Management System fait sur mesure répond aux besoins de multiples contributeurs distants d'avoir une plateforme partagée et multi-langue, permettant d'intervenir dynamiquement dans le contenu. Chaque objet de l'interface sera renseigné suivant le protocole OAI-MH (Open Archives Initiative for Metadata Harvesting), facilitant l'interconnectivité et le partage des données avec d'autres sites sources. Un entrepôt OAI répondra aux demandes des sites tiers pour la récupération du contenu selon un schéma de métadonnées précis (plusieurs candidats seront testés, parmi lesquels Dublin Core).

Dans une optique de valorisation patrimoniale, l'exposition réunira sous une forme numérique des objets fournis à la faveur des partenariats ponctuels avec des institutions culturelles locales ou internationales.

A l'issue de l'exposition pilote que nous souhaitons présenter au colloque DH 2014, le musée sera organisé autour d'expositions thématiques et interdisciplinaires, abordant des questions de "médecine et société". Les expositions feront dialoguer entre eux des points de vue de spécialistes médicaux, d'acteurs du monde de la santé au sens large, mais aussi d'historiens, d'anthropologues ou de représentants des arts et des lettres, afin de faire valoir la complexité sociale et la richesse culturelle des thématiques abordées.

References

- Bates, V., Bleakley G., Goodman S.** (2013), *Medicine, Health and the Arts: Approaches to the Medical Humanities*. Oxon & New York: Routledge.
- Durst, Patrick B, and FAO Regional Office for Asia and the Pacific** (2010). *Forest Insects as Food. Humans Bite Back: Proceedings of a Workshop on Asia-Pacific Resources and*

their Potential for Development, 19-21 February 2008, Chiang Mai, Thailand. Food and Agriculture Organization of the United Nations, Regional Office for Asia and the Pacific.

"QRator" 2011-2013. Consulté le 30 octobre 2013.
www.qrator.org/about-the-project

Ross, C., Gray, S., Warwick, C., Hudson-Smith, A., and Terras, M. (2012). "Engaging the Museum Space: Mobilising Visitor Engagement with Digital Content Creation". *Digital Humanities 2012* [conference paper], July 2012, Hamburg.

"Tales of Things" 2009-2013. Consulté le 30 octobre 2013.
talesofthings.com/about.

Visualization of Historical Knowledge Structures: An Analysis of the Bibliography of Philosophy

Sula, Chris Alen

csula@pratt.edu
Pratt Institute, School of Information & Library S

Dean, Will

wdean@pratt.edu
Pratt Institute, School of Information & Library S

Bibliography is among the oldest forms of knowledge organization, both documenting knowledge artifacts and arranging them into specialized subject categories. Extant bibliographies date back to the 15th century, with more appearing in the 18th and 19th centuries with the rise of trade presses.¹ Though the bibliographic form was eclipsed in the late 20th century by digital systems, print bibliographies offer important insights into the history of knowledge production, especially for works that no longer exist. More importantly, bibliographers' choices of inclusion, exclusion, and subject headings provide unique insight into how works were received in different times and places—and in ways that do not rely on current, anachronistic understandings of disciplines and their topologies.

The potential for datamining these massive bibliographic efforts across five centuries has gone untapped in several senses. First, one might examine statistical patterns in publication dates and topics—akin to the "distance reading" method described by Moretti.² Second, one might analyze geospatial relationships between text and locations, noting when an idea sprouted in a certain area, how long it took to spawn a translation elsewhere, or whether successive versions of the same text signal continued interest in and adoption of certain ideas. Third, one might exploit the associative connections between texts organized under the same subject headings, using those links to examine the intellectual structure of a field, its expansion (or contraction) over time, and its division into new (sub)fields. All of these analyses would be enriched by analyzing multiple bibliographies across different places and times, forming a field of "comparative bibliography" within the digital humanities³.

Philosophy is a prime candidate for this analysis both because extensive research exists on the history of the bibliography of philosophy⁴ and because philosophy once contained nearly all branches of knowledge (the exception being modern science), with other fields splitting off from philosophy from the Renaissance period to the present. General bibliographies of philosophy date back to the 15th century, though Johann Jacob Friesius's 1593 bibliography⁵ may be regarded as the pioneer of the form, which was followed for several centuries by larger compendia. As Jasen's notes, these successive bibliographies reveal shifts in knowledge structures over time: "philosophy has always been in a state of flux, with differing conceptions of scope and interest prevailing at different times. In classical antiquity philosophy encompasses almost all fields of knowledge; but later many of them gradually became separated from philosophy as independent disciplines. Consequently, the history of the subject is largely colored by the continuing tendency to re-define and emphasize special areas

of interest formerly included in larger definitions" ([4] p. 41). A variety of historical classifications is presented in Figure 1.

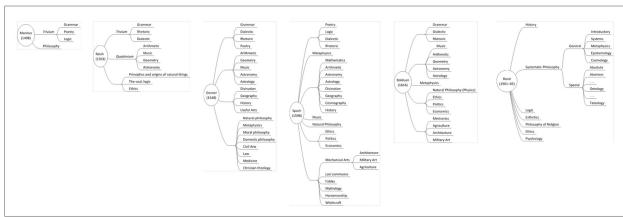


Fig. 1:

Though there is noticeable overlap between these classifications (e.g., the inclusion of the *trivium* and *quadrivium* over several centuries), the differences are also telling, both in terms of the position of subjects within bibliographers' taxonomies and the presence or absence of certain subjects (e.g., medicine, agriculture, law) in one bibliography as compared to others. Jasenas attributes these differences to the influence of past bibliographers, the importance of the subject as perceived by the bibliographer (e.g. Spach's inclusion of ethics), and the university curriculum familiar to each bibliographer—all interesting fodder for a comparative analysis of the field.

The last of these general bibliographies of philosophy came in 1905 with Benjamin Rand's massive compilation⁶, which catalogs over 67,000 books and journals. It has been succeeded only by smaller subject bibliographies. Rand's work drew on past efforts by dozens of bibliographers and contains over 700 subject headings, including 600 entries for specific philosophers' writings and secondary criticism about them. Bynagle notes that "Rand has been criticized, in his own time and since, for omitting certain topics (e.g., philosophy of history and philosophy of language) and for other shortcomings. However, the magnitude of his effort is generally acknowledged and acclaimed, and its product remains even now of some value, particularly for its nearly exhaustive coverage of nineteenth-century authors."⁷

Using Rand's bibliography, this poster presents statistical, geospatial, and network visualizations of the history of philosophy. This data is obtained from a digitized version of Rand's bibliography, which is parsed to extract structured information about each work and its subject heading. The main research questions of the poster are: (1) the emergence of particular topics across time and space (see Figure 2), (2) transmission of ideas as measured through publication location and version information, and (3) the topical shifts in the field over time as reflected by Rand's subject classification of texts.

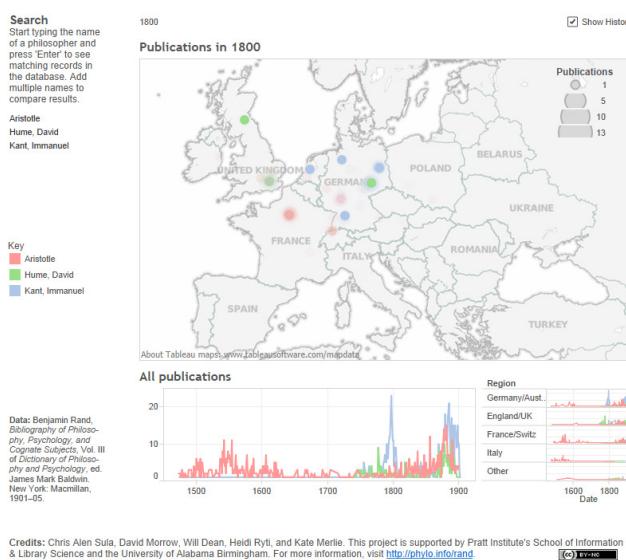


Fig. 2: An interactive geospatial visualization of publications over time, as well as two frequency charts: one showing the total number of

publications over time and another with publications separated by region. Data on Aristotle, Hume, and Kant is displayed.

References

1. Schneider, G. (1934). *Theory and History of Bibliography*. Translated by Ralph Robert Shaw. New York: Scarecrow Press.
2. Moretti, F. (2005). *Graphs, Maps, Trees: Abstract Models for a Literary History*. Verso.
3. For an example in archaeology, see Maximillian Schich, César Hidalgo, Sune Lehmann, and Juyong Park. (2010). "The Network of Subject Co-Popularity in Classical Archaeology, *Bollettino di Archeologia On-line* I, 49–57.
4. Jasenas, Michael. (1973). *A History of the Bibliography of Philosophy*. New York: Georg Olms Verlag Hildesheim.
5. Frisius [Fries], Johann Jacob. *Orationes de officio vitae ministrorum Ecclesiae et de eorumdem concordia*. Tiguri: Apud C. Froeschouerum, 1593.
6. Rand, Benjamin. *Bibliography of Philosophy, Psychology, and Cognate Subjects*. Vol. III of *Dictionary of Philosophy and Psychology*. Edited by James Mark Baldwin. New York: Macmillan, 1901–05
7. Hans E. Bynagle (1997). *Philosophy: A Guide to the Reference Literature*, 2nd ed. Englewood, Colo. Libraries Unlimited, Inc. pp. 87–88.

TextGrid: Creating, archiving, publishing and exploring digital editions and other humanistic research data via a Virtual Research Environment

Söring, Sibylle

sibylle.soering@sub.uni-goettingen.de
Georg-August-Universität Göttingen, Niedersächsische Staats- und Universitätsbibliothek, Germany

Veentjer, Ubbo

Georg-August-Universität Göttingen, Niedersächsische Staats- und Universitätsbibliothek, Germany

Funk, Stefan

Georg-August-Universität Göttingen, Niedersächsische Staats- und Universitätsbibliothek, Germany

Today, scholarly digital editions represent one of the core application areas of the Digital Humanities. Virtual Research Environments (VREs) enable and support the creation, publication, and long-term archiving of such data. Answering an increasing demand for digital and collective research features in the humanities, the joint project TextGrid, funded by the German Federal Ministry of Education and Research has, since its start in 2006 and in continuous exchange with the research community, developed a VRE that aims at mapping the entire research process.

Whereas the **TextGrid Laboratory** (TextGridLab) contains a versatile open source software for editing and generating digital sources collaboratively in a protected virtual environment, and allowing for a differentiated user rights management, the **TextGrid Repository** (TextGridRep) offers an open, XML/TEI-based long-term research archive, in which both the text and image data generated with the TextGridLab, as well as external digital objects, can be published, browsed, explored, analysed, cited, and archived.

A crucial factor of a VRE's success and impact is the use of both technological and semantic standards. The TextGrid architecture supports, amongst common metadata standards, the markup language XML together with the well-established markup data format TEI. They reflect international standards for the sustainable, searchable and reusable mark-up of humanistic sources, especially of digital editions. A differentiated user rights management facilitates collaborative work on a shared

project in a non-public environment. Tools, data, and methods can be used mutually, regardless of the operative system, software equipment, or location.

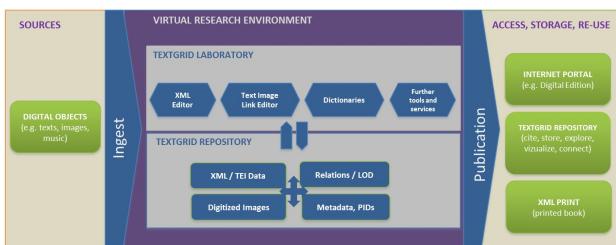
In addition to the tools and services available in the TextGridLab, the TextGridRep provides the user with the possibility to save, publish, and search a variety of digital resources such as XML/TEI encoded texts, images and databases, therefore supporting the creation of Linked Open Data. Thus, both core components of the VRE, the TextGridLab and the TextGridRep, are aligned for an optimal interaction, interlinkage, and workflow between and with one another.

Beyond the creation of resources, TextGrid ensures the persistent availability of and access to research data as well as optimal interconnectivity, supporting international standards. Collaborative research is facilitated by e.g. the annotation of images; further annotation features are currently evaluated.

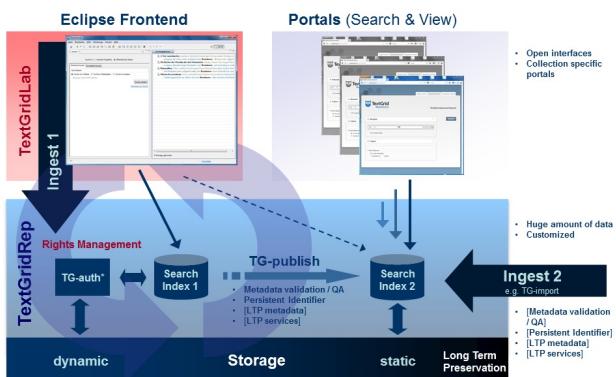
Thus, TextGrid facilitates the creation of digital editions, from the provision and creation of primary data in XML to published and citable research data; data that can also be made available in an external online portal, or be exported into a print-ready document (PDF).

As of today, TextGrid has approx. 1.500 registered users and approx. 40 research projects from a broad spectrum of humanistic disciplines ranging from philology, cultural studies, medieval studies, jewish and ecclesiastical history to linguistic and musicological studies. Amongst them are single scholars and medium-scale research groups such as "Theodor Fontane's notebooks", as well as large, long-term projects like "Johann Friedrich Blumenbach – online" or "IDIOM – Interdisciplinary Dictionary of Classic Mayan".

The poster will display how a complete scholarly workflow can be mapped via a VRE like TextGrid – from collecting and generating primary data through enriching it with metadata and XML/TEI, and finally publishing it in a portal and/or a repository, following sustainable standards and thus allowing for citation, long-term accessibility, further interlinking and scholarly reuse. In this scope, the poster will not only focus on text-based research data, but will also explore technologies allowing for a web based annotation, viewing and publishing of image formats such as Digitilib.



TextGrid Repository Architecture



References

<http://www.textgrid.de/en/>

<https://www.textgrid.de/en/registrationdownload/download-and-installation/>
<http://www.textgridrep.de/>
<http://www.tei-c.org/index.xml>
<http://www.uni-goettingen.de/en/303691.html>
<http://www.blumenbach-online.de/index.php?id=2&L=1>
<http://www.iae.uni-bonn.de/forschung/forschungsprojekte/laufende-projekte/idiom-dictionary-of-classic-mayan/idiom-english-project-description>
<http://digilib.berlios.de/>

The Arabic Papyrology Database

Thomann, Johannes

johannes.thomann@aoi.uzh.ch
University of Zurich

1. Peculiarities of the Arabic Writing System

The Arabic writing system in general and the writing conventions in papyri in particular require special display formats. As in most Afro-Asiatic writing systems, short vowels are not represented by letters, but there exist vocalisation marks (harakāt), written above or below the letters. Further, one-letter-words and the article are written together with the following word and the clitical pronouns are attached to the preceding word. In transliteration these words are separated by a hyphen. Finally, 15 letters (al-ḥurūf al-mu'jamah) are distinguished by diacritical pointing from letters with the same basic form¹. However, these dots are only occasionally found in early Arabic documents. Most modern editions of early and classical Arabic texts do not account for these peculiarities of the Arabic writing system and normalize the texts according to modern orthographical rules without using vowel-signs. While this established practice may be appropriate for literary texts, a more elaborate procedure is needed for documentary texts.

2. The Data

There are two main groups of premodern Arabic documents: Inscriptions and papyri in the broader sense. The second group consists of about 150'000 documents written on different material such as papyrus, parchment or paper during the period from the 7th to the 16th centuries, of which about 2'500 have been edited². Another 10'000 unpublished documents are described or mentioned in papyrological publications. All these documents provide information on almost every aspect of Islamic history. Despite their importance, they still do not receive a high level of attention in historical research³.

3. Methodology

Two leading ideas were at the beginning of the Arabic Papyrology Database (APD) [www.ori.uzh.ch/apd]: On one hand it should be a research tool which makes metadata and full texts of all edited documents easily accessible, and on the other it should overcome the limitations of printed editions. As already mentioned, modern editions present the texts in a normalized form according to modern Arabic orthography. Some high quality editions describe the diacritical pointing and vocalisation marks in the critical apparatus, and a limited number of editions provide word indices in transliteration.

For the APD a completely new approach of organizing Arabic text was developed. Instead of the single text level approach used in print and in other database projects, the APD presents texts in five levels. This approach is unprecedented in the entire field of Arabic studies. On the first level only diacritical dots found in the document are written and all observations of gaps, deleted texts and redundant words are indicated by sigla and brackets [orientw.uzh.ch:8080/apd/requisits3c.jsp].

On the second level these marks are removed and the text is broken into single words. On the third level missing diacritical dots are added, as it is common in text editions. On the fourth level vowel marks are added in providing a full phonological representation. On the fifth level the text is latinised with segmentation of the elements written together in Arabic script [orientw.uzh.ch:8080/apd/requisites2.jsp]. Further, each element of the fifth level is connected to a lexicon and a list of grammatical forms. The levels of text are hierarchically organized, and variant readings and remarks are attached to their appropriate level. Search is not only possible along each level, but orthogonal searches across levels for the corresponding or neighbouring word in another level can be carried out as well.

4. State of the Project

The APD was initiated by Andreas Kaplony and Johannes Thomann at Zurich University in 2004. Today it is a joint project of the Universities of Zurich, Munich (LMU) and Vienna. The APD was online and freely accessible from its beginning. A PhD project by one of its collaborators was based on the APD⁴, and another research project would have been impossible without the advanced search capabilities across text levels in the APD⁵. At present, 1563 full text documents are available, and during the next two years the remaining edited documents will be entered into the APD. There are mutual references in the APD and the Trismegistos database, and soon texts of the APD will be imported by the Papyrus Navigator, made possible by the XML export engine of the APD in EpiDoc format.

References

1. Gacek, A. (2009). *Arabic Manuscripts: A Vademecum for Readers*. Leiden: Brill.
2. Sijpesteijn, P. M. (2005). *Checklist of Arabic Papyri*. Bulletin of the American Society of Papyrologists. 42: 127–166. Updated version (last accessed 1 November 2013): www.aoi.uzh.ch/islamwissenschaft/forschung/isap/isapchecklist.html
3. Sijpesteijn, P. M. (2009). *Arabic Papyri and Islamic Egypt*. The Oxford Handbook of Papyrology. Oxford: University Press, pp. 452–472.
4. Grob, E. M. (2010). *Documentary Arabic Private and Business Letters on Papyrus: Form and Function, Content and Context*. Berlin: de Gruyter.
5. Kaplony, A. (2008). *What Are Those Few Dots For?* Thoughts on the Orthography of the Qurra Papyri (709–710), the Khurasan Parchments (755–777) and the Inscription of the Jerusalem Dome of the Rock (692). *Arabica* 55 (1): 91–112.

“Crowdsourcing Annotation and the ‘Social Edition’: Ossian Online.”

Tonra, Justin

justin.tonra@nuigalway.ie
National University of Ireland, Galway

Barr, Rebecca

rebecca.barr@nuigalway.ie
National University of Ireland, Galway

1. Introduction

1.1. Overview

James Macpherson's *Ossian* poems were the international sensation of the eighteenth-century. First published in 1760, Macpherson's work caused a literary furore. Ostensibly translations from Gaelic manuscripts, the poems were

published as fragments of a lost Celtic epic, salvaged from a dying oral culture and translated for the edification of a modern readership. Despite the controversial provenance of the *Ossian* poems, they transformed European literature; their impact was profound, international and long lasting, initiating the Romantic movement in Ireland, Britain, Europe, and beyond.

1.2. Methodology

Ossian Online is a new initiative to freshly edit and make available this profoundly influential work of eighteenth- and nineteenth-century European culture. It will work on the principle of the ‘collaboratory’: providing an online infrastructure for scholarly collaboration. As a platform in which participants can annotate, debate, and engage, this project will create an innovative space for interdisciplinary dialogue, where scholarly debate and exchange can occur in real-time.

2. Proposal

The past five years have witnessed an exponential growth in the use of social media for scholarship and communication in eighteenth-century studies (*Eighteenth-Century Questions*¹, *The 18th-Century Common*², *18thConnect*³, *Mapping the Republic of Letters*⁴). *Ossian Online* harnesses this critical mass and directs its potential towards the online scholarly edition. By creating a new online edition of the poems which visualises textual variation, evolution, and genetic relations, and altering the medium in which the text is presented, this project will bring *Ossian* to a global audience.

Ossian Online will also act as a test case for new approaches to humanities research, bringing greater immediacy and interdisciplinarity to the fundamental practices of academic communication than are afforded by traditional models of scholarly publication. The rewards of this endeavour will be apparent not just in the synthesis of different disciplinary insights, but in the challenges it poses to established disciplinary conventions. *Ossian Online* uses social media technologies to crowdsource annotations to a new edition of the *Ossian* poems. The project closely follows many of the recent articulations of the possibilities of the ‘social edition’,^{5 6}. It also provides a practical example of an edition which enacts one of the many potential affordances of social media for scholarly editing and annotation. *Ossian Online* aims to contribute to the description of an active typology of the emergent ‘social edition,’ which remains more theorised than practiced. More broadly, this paper will seek to “extend our understanding of the scholarly edition in light of new models of edition production that embrace social networking and its commensurate tools.⁷

The multidisciplinary appeal of *Ossian* makes it an ideal candidate to test a set of technologies which promise to use participatory experience to reorient the role of the scholarly editor “away from that of ultimate authority and more toward that of facilitator of reader involvement”⁸. Scholars from the range of disciplines that study *Ossian* (literature, history, Irish studies, Scottish studies, Celtic studies, romanticism, textual studies, book history) are a crowd—as McGann has put it—“who have yet to be sourced”⁹. To date, crowdsourcing has been used for different scholarly ends (including transcription, correction, and identification of data), but this represents one of the first occasions on which the wisdom of the crowd will be leveraged to critically annotate a literary work. Building on the principles of existing crowdsourcing software (*Transcribe Bentham*¹⁰, *Candidate 2.0*¹¹, *Prism*, *CommentPress*¹³, *Digress.it*¹⁴), *Ossian Online* will develop an interface for the collaborative research environment that will satisfy the particular needs of the literary text and reinvigorate related scholarship. Moving *Ossian* online preserves the core-values of the humanities while articulating them through new opportunities offered by the digital revolution. It will facilitate a forum in which multiple scholarly perspectives can be synthesised, through an interdisciplinary research environment.

Interest in the ‘social edition’ is growing within scholarly editing and digital humanities communities. In a similar manner

to the recent 'Social, Digital, Scholarly Editing' conference at the University of Saskatchewan¹⁵, this paper will address the theoretical, practical, and social effects of the collaborative editorial possibilities enabled by the development of digital platforms.

This paper will have two particular focuses: first, to provide a critique of social media platforms and technologies used by *Ossian Online*, and suggest which are best suited to fulfilling the needs of 'social edition' developers. Second, it will articulate the current possibilities and challenges of constructing a 'social edition,' outlining future directions for "the organization of digital text [...] to promote social interaction within and around it"¹⁶.

References

1. **Eighteenth-Century Questions.** Facebook. Web. 17 Oct. 2013.
2. **The 18th-Century Common.** Wake Forest University. Web. 17 Oct 2013.
3. **18thConnect.** Texas A&M University. Web. 17 Oct 2013.
4. **Mapping the Republic of Letters.** Stanford University. Web. 17 Oct. 2013.
5. **Siemens, Ray et al.** *Pertinent Discussions Toward Modeling the Social Edition: Annotated Bibliographies*. Digital Humanities Quarterly 6.1 (2012): n. pag. Web. 17 Oct. 2013.
6. **Siemens, Ray et al.** *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 (2012): 445–461. llc.oxfordjournals.org. Web. 17 Oct. 2013.
7. **Siemens, Ray et al.** *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 (2012): 445–461. llc.oxfordjournals.org. Web. 17 Oct. 2013. 447.
8. **Siemens, Ray et al.** *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 (2012): 445–461. llc.oxfordjournals.org. Web. 17 Oct. 2013. 446.
9. **McGann, Jerome.** *Introduction*. Online Humanities Scholarship: The Shape of Things to Come, ed. Jerome McGann. Houston: Rice University Press, 2010. 1-4. Print. 2.
10. **Transcribe Bentham.** University College London. Web. 17 Oct. 2013.
11. **Candide 2.0.** New York Public Library. Web. 17 Oct. 2013.
12. **Prism.** Scholars' Lab, University of Virginia. Web. 17 Oct. 2013.
13. **CommentPress.** Institute for the Future of the Book. Web. 17 Oct. 2013.
14. **Digress.it.** Institute for the Future of the Book. Web. 17 Oct. 2013.
15. **Social, Digital, Scholarly Editing.** University of Saskatchewan. 11-13 Jul. 2013.
16. **Fitzpatrick, Kathleen.** *CommentPress: New (Social) Structures for New (Networked) Texts*. Journal of Electronic Publishing 10.3 (2007): n. pag. Web. 17 Oct. 2013.

A Quantitative Analysis for the Authorship of Saikaku's Posthumous Works in the Seventeenth Century

Uesaka, Ayaka

ayaka.u26@gmail.com
Doshisha University, Japan

Murakami, Uesaka

mamuraka@mail.doshisha.ac.jp
Doshisha University, Japan

I. Introduction

In this paper, we focus on Saikaku Ihara's posthumous works collections, especially *Yorozu no humihougu* (万の文反古) (*A Scrapbook of Old Letters*; 1696). Saikaku Ihara (井原西鶴) is one of the most famous writers of the Edo period (1603– 1868) in Japan. It is said that he wrote 23 works in 10 years. However, that achievement has not been fully verified.

Identifying authors of novels in this era is difficult, because typically their authors did not sign the books. In addition, Saikaku's posthumous works were edited and published from 1693 to 1699 by his student Dansui Houyou(北条団水). It became more difficult to identify the authors of Saikaku's posthumous works.

In our study, we focused on *Yorozu no humihougu* because many Saikaku's researchers have raised questions about the authorship. Saikaku researchers have tried to identify his works by investigating their history, content, format and so on. However, it remains unclear which works are really by Saikaku. Meanwhile, the potential of quantitative analysis of textual data and the related field of the digital humanities have also dramatically advanced. For this reason, this study verifies the text of *Yorozu no humihougu* using a quantitative approach.

II. Database of Saikaku's works

To resolve the Saikaku authorship problem, we made a database of his works with Saikaku researchers, so this database is the only one on Saikaku's works presently and has a high degree of reliability. We made this database based on Shinpen Saikaku Zenshu (新編西鶴全集) (Figure 1)[1].

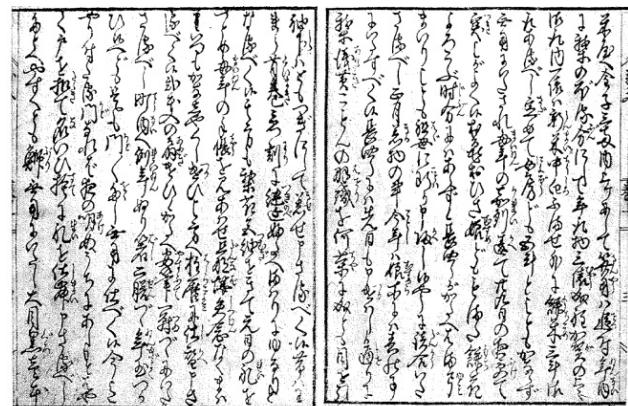


Fig. 1: 17th century publication

Table.1 is a part of the database from Saikaku's works used for this analysis. All sentences were divided into words. Moreover, information was added as required for the analysis. According to our database, the total words in 23 Saikaku works are about 578,617 words.

Work	Volume	Words	Part of speech	Other information
万古	卷一	世帯	名詞	セタイ せたい せたい
万古	卷一	の	助詞	△ の の
万古	卷一	大事	名詞	ダイジ だいじ だいじ
万古	卷一	は	助詞	△ は は
万古	卷一	正月仕舞	名詞	△ シュウガツミ しゅうがつじまい しゅうがつじまい
万古	卷一	十二月九日	名詞	△ ジュウニンヒル じゅうにんがつここのか じゅうにんがつここのか
万古	卷一	の	助詞	△ の の
万古	卷一	書中	名詞	ショチウ しゅちゅう しゅちゅう
万古	卷一	句読点	△	,
万古	卷一	伊勢屋十左衛門	名詞	イセヤジロウザエモン いせやじろうざえもん
万古	卷一	船	名詞	フネ ふね ふね
万古	卷一	、	句読点	△ , ,
万古	卷一	十二日	名詞	△ ジュウニンヒル じゅうににち じゅうににち
万古	卷一	に	助詞	△ に に
万古	卷一	くだりにつき	動詞	連用 △ くだりにつく くだりにつく

Fig. 2: Database of Saikaku's works

III. Analysis and results

In general, Saikaku's works are made up of many short stories (chapters), so we used information of each chapter in our analysis. In addition, we used four posthumous works other than *Yorozu no humihougu* as one group (Table 2). Then, we compared *Yorozu no humihougu* to four other posthumous works.

Works name	Chapter
<i>Yorozu no humihougu</i> (万の文反古)	17 chapters
<i>Four other posthumous works</i>	<i>Saikaku oki miyage</i> (西鶴置土産)
	15 chapters
	<i>Saikaku oridome</i> (西鶴織留)
	23 chapters
	<i>Saikaku zoku turezure</i> (西鶴俗つれづれ)
	18 chapters
	<i>Saikaku nagori no tomo</i> (西鶴名残の友)
	27 chapters

Fig. 3: Saikaku's posthumous works

At first, we examined the appearance rate of the seven principal parts of speech: nouns, particles, verbs, auxiliary verbs, adjectives, adverbs and adnominal adjectives. Figure 2 shows the results of the analysis on the appearance rate, using the principal component analysis (PCA) with a correlation matrix. The horizontal axis shows the importance of the first principal component, and the vertical axis shows the second. In this figure, indicating differences revealed by PCA, 17 chapters of *Yorozu no humihougu* is on the right and 83 chapters of *four other posthumous works* on the left.

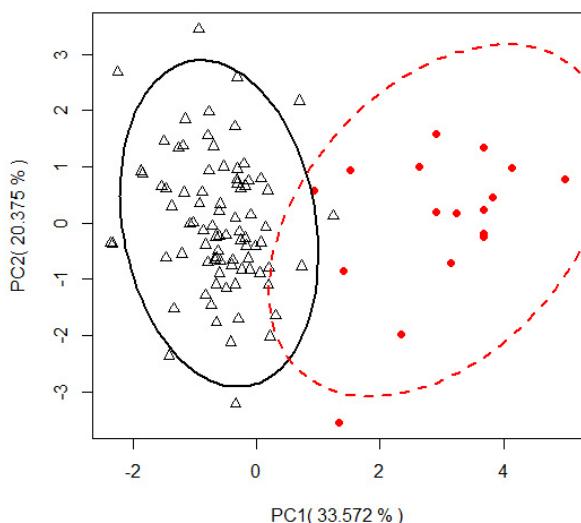


Fig. 4: CA results for *Yorozu no humihougu* (万の文反古) posthumous works (These circles drawn on the figure are 95% confidence ellipse)

Upon examining the first principal component, we found that differences in the use of particles and verbs (Table 3). In *Yorozu no humihougu*, the appearance rate of verbs is higher and those of the particles are lower, compared with *four other posthumous works*.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
nouns	-0.16	0.70	0.25	-0.06	0.51	0.10	0.39
particles	-0.54	-0.04	-0.19	0.28	-0.44	0.38	0.51
verbs	0.62	-0.07	0.12	-0.03	-0.25	-0.25	0.69
auxiliary verbs	-0.28	-0.60	0.02	0.20	0.58	-0.31	0.29
adjectives	-0.25	0.15	-0.58	-0.58	-0.07	-0.46	0.12
adverbs	0.37	-0.12	-0.59	-0.08	0.37	0.59	0.11
adnominal adjectives	-0.13	-0.34	0.44	-0.73	-0.03	0.36	0.10
Proportion of Variance	33.57	20.37	16.07	13.74	8.9	7.04	0.3
Cumulative Proportion	33.57	53.94	70.01	83.75	92.65	99.69	100

Fig. 5: The result of PCA

Furthermore we examined by Welch's t-test at the 0.05 significance level (Table 4). Significant differences were found in verbs and particles.

	p-value	t-value
nouns	0.07717	1.8599
particles	3.217E-08	9.2974
verbs	6.225E-11	-13.1611
auxiliary verbs	0.001555	3.6522
adjectives	0.06828	1.9167
adverbs	0.001407	-3.6964
adnominal adjectives	0.3514	0.9494

Fig. 6: The result of Welch's t-test. (Note: a positive t-value indicates that the word is characteristically used in four other posthumous works)

Next, we examined the words of verbs and particles characteristically used in each work by using Welch's t-test for determining whether there exists a difference between the averages of *Yorozu no humihougu* and four other posthumous works. Of the 1,625 types (15,113 words) of verbs, Welch's t-test was performed on each of the 47 types (8,055 words) that appeared more than 51 times. Similarly, of the 53 types (29,994 words) of particles, Welch's t-test was performed on each of the 31 types (29,842 words). We examined 10 words with the smallest p-values.

Table 5 shows that "sourou," "mousu," and "gozasourou" are the most frequently used words of verbs in *Yorozu no humihougu*, and "su," "iu," "naru," "ari," "yuku," "sumu," and "tamau" are in *four other posthumous works*. From these results, honorific words of verbs are assumed to be characteristically used in *Yorozu no humihougu*.

	p-value	t-value
su (す)	2.2.E-16	13.40
iu (いう)	2.2.E-16	13.70
sourou (そうろう)	1.2.E-12	-18.82
mousu (もうす)	1.9.E-10	-13.20
naru (なる)	2.1.E-10	7.66
ari (あり)	6.3.E-08	6.81
gozasourou (ござそうろう)	2.9.E-06	-6.98
yuku (ゆく)	7.3.E-06	4.87
sumu (すむ)	7.3.E-06	4.74
tamau (たまう)	9.5.E-06	4.67

Fig. 7: The result of Welch's t-test. (Note: a positive t-value indicates that the word is characteristically used in four other posthumous works)

Table 6 shows that "he," "bakari," "ni," "nite," and "ha" are the most frequently used words of particles in *Yorozu no*

humihougu, and "zo," "te," "do," "kasha," and "tote" are in *four other posthumous works*.

	p-value	t-value
zo (ぞ)	5.2.E-09	6.97
te (て)	7.1.E-06	5.83
do (ど)	3.0.E-05	4.42
kashi (かし)	5.3.E-04	3.83
tote (とて)	8.6.E-04	3.6
he (へ)	1.0.E-03	-3.92
bakari (ばかり)	3.6.E-03	-3.25
ni (に)	4.4.E-03	-3.09
nite (にて)	5.2.E-03	-3.17
ha (は)	8.5.E-03	-2.93

Fig. 8: The result of Welch's t-test. (Note: a positive t-value indicates that the word is characteristically used in four other posthumous works)

IV. Conclusion

We conducted the comparative analysis among *Yorozu no humihougu* and *four other posthumous works* using a quantitative approach. *Yorozu no humihougu* was revealed to be characterized as having a higher the appearance rate of verbs and a lower the appearance rate of particles than in four other posthumous works. Furthermore, we analyzed the words of verbs and particles characteristically used in each work by using Welch's t-test, in *Yorozu no humihougu*, honorific verbs are used more frequently than in *four other posthumous works*, while the particles "he," "bakari," "ni," "nite," and "ha" appear more often. These results indicate that *Yorozu no humihougu* and *four other posthumous works* are quite different in the appearance rates of parts of speech as well as the words of verbs and particles.

Among the works created by Saikaku, *Yorozu no humihougu* is the only work written in an epistolary style. Analyses on the characteristics of the epistolary style of writing should be performed in the future in order to clarify any doubt concerning the author of *Yorozu no humihougu*.

In addition to the appearance rate of parts of speech, other information such as the appearance rate of words, as well as the works of Dansui Houjyou, should be analyzed.

References

- Noboru Asai et al. (2000), "Shinpen Saikaku Zenshu (新編西鶴全集)" vol.1–4, Benseishupan
- Munemasa isoo (1969), "Kanazoushi kara Ukiyozoushi he (仮名草子から浮世草子へ)", Shibundou
- Yamaguchi takeshi (1929), "Saikaku meisakushu ge (西鶴名作集下)", Nihonmeityo zenshu kankoukai
- Teruoka yasutaka (1953) "Saikaku kenkyu note (西鶴研究ノート)", Tyuuoukouronsha
- Taniwaki masachika (1981) "Saikaku kenkyu ronkou (西鶴研究論叢)", Sintensha
- Nakamura yukihiko (1982), "Nakamura yukihiko cyogitsushu (中村幸彦著述集)", Tyuuoukouronsha

The Proportional Sizes of Genres in Eighteenth- and Nineteenth-Century English-Language Books

Underwood, Ted
tunder@illinois.edu
University of Illinois

This poster represents the first stage of a larger project on automated genre classification in a collection of a million

volumes (the HathiTrust collection of English-language books 1700–1950). Much existing work on genre classification has focused on fine distinctions between subgenres of fiction¹ or poetry.² But in mapping a large digital library, "genre" is a term that has a range of different meanings appropriate to different scales of analysis.³ Before we can even attempt to make subtle discriminations between, say, "the sensation novel" and "detective fiction," we need to create a simpler map of the collection that identifies sections of each volume broadly as "prose fiction" or "drama," or for that matter as "publishers' ads" or a "library bookplate." At the University of Illinois, we've developed an automated workflow that we feel does this initial mapping accurately enough for distant reading. The technique was described at the IEEE Big Humanities workshop in October, 2013,⁴ with a brief illustration, but we haven't yet presented results across a broad range of genres. That's what this poster will do.

In an ideal world, structural features of a volume would be coded manually with TEI. But since large digital libraries collect plain text rather than TEI, mining large collections will initially require an automated strategy. Our strategy involves training an ensemble of classifiers to recognize genres and aspects of volume structure at the page level. For instance, we train classifiers to recognize "prose fiction" and "drama," but also "tables of contents," "bookplates," "date due slips," and "publishers' ads." By themselves these classifiers can achieve reasonable accuracy, but we've also found it useful to pair them with another level of machine learning: a hidden Markov model trained on page sequences that implicitly learns about the larger-scale patterns that organize page-level features into volumes. (For instance, indexes are more likely to follow nonfiction than fiction, and not at all likely to precede fiction.)

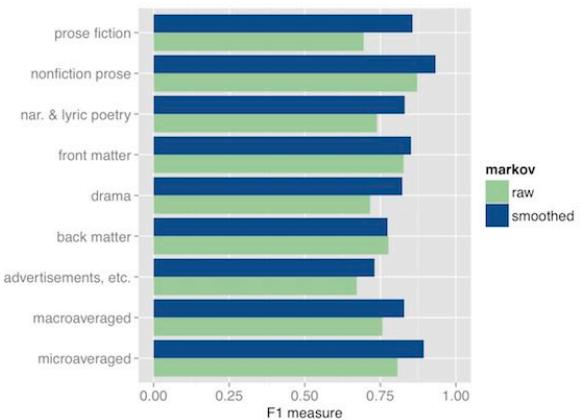


Fig. 1: Tenfold cross-validation of page-level classification. The top seven rows are F1 measures for individual genres; the bottom two rows reflect macro- and micro- averaged F1 measures for all genres. Green bars indicate raw classification accuracy before smoothing; blue bars reflect gains from hidden Markov smoothing.

A preliminary tenfold cross-validation of this technique is presented in Figure 1. This was based on relatively modest training data (101 volumes); by the time we present in Lausanne we expect to be able to increase the size of the training set by an order of magnitude, and significantly increase accuracy. But even with an F1 metric in the range of 85–90%, the technique is accurate enough to illuminate the broad outlines of book history, revealing roughly what proportion of the collection is devoted to nonfiction, or fiction, or (as illustrated in Fig. 2) publisher's advertisements. Here we've focused specifically on publishers' advertisements in volumes of fiction, and graphed their prevalence as a percentage of words in the fiction corpus.

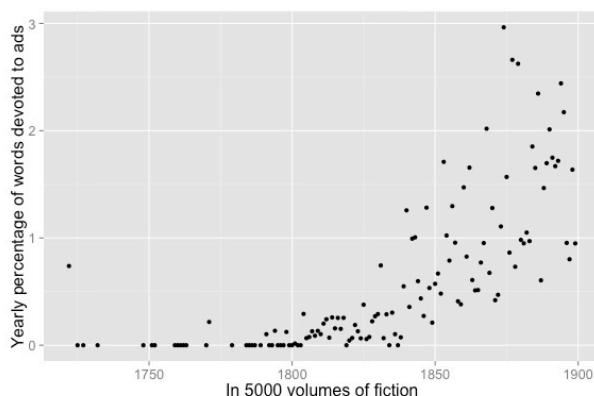


Fig. 2: The yearly percentage of words devoted to publishers' advertisements, in 5000 volumes of fiction selected randomly from a larger corpus of 32,200.

In the poster we will include a streamgraph visualizing the relative sizes of major literary genres across time (for instance, verse drama, lyric poetry, narrative poetry, prose fiction), as well as smaller graphs that visualize the history of particular structural features within volumes (for instance, for the prose footnotes that occupy a great deal of space in eighteenth- and nineteenth-century volumes of poetry).⁵

References

1. S. Allison, R. Heuser, M. Jockers, F. Moretti and M. Witmore. (2011) *Quantitative Formalism: An Experiment, Stanford Literary Lab Pamphlet Series*. [Online]. Available: litlab.stanford.edu/?page_id=255
2. B. Yu. (2008) *An Evaluation of Text Classification Methods for Literary Study*, in *Literary and Linguistic Computing*, Vol. 23 (2008): 327-343.
3. M. Santini. (2004) *State-of-the-Art on Automatic Genre Identification*, Information Technology Research Institute Technical Report Series, ITRI, University of Brighton, Jan. 2004.
4. T. Underwood, M. L. Black, L. Auvil, and B. Capitanu (2013). *Mapping Mutable Genres in Structurally Complex Volumes*. Proceedings of IEEE Big Data 2013. arxiv.org/abs/1309.3323
5. This poster reflects work by Shawn Ballard, a graduate student in English at the University of Illinois, who may be added as co-author in the final version. The project has also been supported by the Andrew W. Mellon Foundation, the National Endowment for the Humanities, and the American Council of Learned Societies.

Kiln: XML Publishing Framework

Vieira, Miguel

King's College London, United Kingdom

Norrish, Jamie

King's College London, United Kingdom

Getting an XML-based website up and running can be a repetitive, time consuming and tedious task. It usually also requires someone with good technical knowledge to bring all of the pieces together into a complete package. Kiln is a tool that allows a non-technical user to quickly and easily create a working web application that provides base functionality for publishing, indexing and searching XML source content.

The user is free to focus on developing the content, and can easily assess the way in which content is used and review changes to see how they are reflected in the published output. Importantly this ease of use does not preclude the customisation of the existing functionality, nor the addition of new tools, processes and outputs.

Kiln, previously known as xMod, is an open source multi-platform framework for building and deploying complex websites whose source content is primarily in TEI/XML. It brings together various independent software components into an integrated whole that provides the infrastructure and base functionality for such sites.

Kiln is developed and maintained by a team at the Department of Digital Humanities, King's College London. Over the past years and over several versions, Kiln has been used to generate more than 50 websites for digital humanities research projects which have very different source materials and customised functionality.

The main Kiln components are:

- Apache Cocoon web development framework for XML processing.
- Jetty web application server.
- Apache Solr, a searching platform for indexing, searching and browsing of contents.
- Sesame, a framework for processing RDF data.

In a production web server context, Kiln integrates with other web publishing tools to support images, maps and other data sources, like relational databases. It can equally easily be used to provide a full web application solution, or as a backend providing content to be surfaced by some other technology, such as Django or WordPress.

Kiln has been developed around the concept of the separation of roles, allowing people with different backgrounds, knowledge and skills to work simultaneously on the same project without interfering with one another's work. The parts of the system used by developers, designers and content editors are distinct. For example, the templating engine handles the general structure of an output document in individual files, with reference to separate sources which supply the individual content elements. Templating is also achieved using a lightweight syntax that allows frontend development to take place without any knowledge of XSLT. Kiln has two competing design goals: to support the development of unique, complex web applications; and to provide an out-of-the-box system suitable for a single non-technical person to publish a TEI-based site. The former demands not only the customisability of every component, but also the avoidance of any technical magic that makes one way easier at the cost of another way being harder. The latter requires a large amount of built-in behaviour that can be easily tweaked in isolation, and excellent documentation. Kiln's documentation includes a tutorial showing how to customise each of the major elements of a site, as required beyond the provided defaults.

In comparison with other publishing tools, such as XTF, SADE and TE-ICHI, Kiln offers some advantages. It is mature and flexible, as it has been in active development for 10 years since its initial version, and has been used in many research projects. It is standalone (beyond an installation of the Java language), requiring no other software for any of its functionality, and it provides a working site, including a faceted search, with no more than placing TEI files in a particular directory.

The proposed poster will provide a diagrammatic overview of a Kiln project structure, and an accompanying interactive demo will show the system in action, from creating a new project through to displaying content as a website.

References

1. kcl-ddh.github.io/kiln/
2. A list of some of these projects is available at kiln.readthedocs.org/en/latest/projects.html.
3. cocoon.apache.org/2.1/
4. www.eclipse.org/jetty/
5. lucene.apache.org/solr/
6. www.openrdf.org/
7. xtf.cldlib.org/
8. www.bbaw.de/telota/software/sade/sade-1
9. www.teichi.org/

An ontology for 3D visualisation of cultural heritage

Vitale, Valeria

valeria.vitale@kcl.ac.uk
King's College London, UK

To date, 3D computer graphics and modelling techniques have been used in the study of the ancient world mainly as a means to display traditional research. The value of these digital techniques has been often assessed merely on the degree of graphic aesthetic quality (Favro 2006).

The pursuit of photorealism has proven ineffective in engaging the audience (Champion and Dave 2007) but also misleading, as it suggests that is possible to reproduce an artefact or a scene «exactly as it was» in the past (Baker 2012). Behind every scholarly 3D visualisation is a thorough study of excavation records, iconographic documentation, ancient literary sources, artistic canons and precedents (Hermon 2008). However, this valuable research is not always detectable in the final visual outcome.

*The London Charter for the Computer-based Visualisation of Cultural Heritage*¹ made a huge step forward in the regulation of scholarly 3D visualisation—prescribing that researchers' choices and motivation must all be documented. No 3D model can be considered a scholarly resource if its research method is not «transparent» (Forte 2012). The London Charter presents methodological guidelines for recording this data, but does not go as far as to offer a formal framework in which to place this information; each modeller is left to simply follow their own style. Moreover, the clients who commissioned the 3D model (such as museums or other cultural institutions) are frequently more interested in the final product than in its rationale, which is often completely overlooked and not circulated (or dropped from the budget line altogether). Time and resource constraints not only affect the accuracy and availability of the documentation, but also make it very unlikely that a researcher, or even a team, develops more than one visualisation of the same cultural heritage place/object, perpetuating the naive idea that only one visualisation is possible or correct.

The growing compatibility between 3D content and web browsers² allows the use of RDF technology to, potentially, connect the 3D model and its parts, internally with each other—identifying and defining relationships—and externally with online information about the material remains, previous publications, primary and secondary sources, and with available alternative visualisations of the same object (that share the same controlled vocabulary).

Ontologies for cultural heritage are already commonly used in the management of museum collections and databases³. However, they tend to focus on material artefacts and to meet the specific needs of museum curators and cataloguers. Therefore, they do not seem the most suitable means to deal with digital objects (that are hypothetical representations of material objects), to state methodological relationships, or describe a scholarly process.

The proposed ontology for 3D visualisation for cultural heritage will:

- Describe the 3D digital object. After assigning a specific URI to each element of a 3D digital visualisation (@prefix "obj": <http://hypothetical3dontology.kcl.ac.uk/objects/>), they will be associated to metadata (such as creator(s), software(s) used, formats available) and to a formal description of the cultural heritage object they represent.

For example obj:001 rdf:type art:shaft. Where @prefix "art": <http://hypothetical3dontology.kcl.ac.uk/ArtVocabs/>⁴

- Describe the 3D digital object's relationships with other 3D digital objects⁵. Through a dedicated namespace (@prefix "tdvo": <http://hypothetical3dontology.kcl.ac.uk/threedisontology/>) it will be possible to state and describe properties, values and relationships of the 3D digital objects such as

- the relationship between a 3D digital object and the file it belongs to (obj:001 tdvo:isPartof obj:3Dfile.max);

- the relationships between different objects within the same file (obj:001 rdf:type art:shaft. obj:010 rdf:type art:column. obj:001 tdvo:isPartof obj:010).

- Describe 3D digital object's relationships with its physical referent. Through a digital geographical gazetteer such as Pleiades⁶, the 3D digital object will be linked to the place where the visualised building (or other referent) is located. For example, for a file "3Dfile.max" visualising the Odeon in Aphrodisias, we will have: 3Dfile.max gawd:depicts pleiades:638753/odeon. Different 3D visualisations could be connected to the physical building and be available alongside the photographic documentation linked to Pleiades via Flickr.

- Assess and represent the level of speculation involved in the creation of each element, presenting 3D visualisation more as a scientific hypothesis than an «exact reconstruction». For example obj:001 tdvo:hasCertainty tdvo:certainity⁸

where the level of certainty from 6 (maximum) to 0 (minimum) would be defined as follow:

tdvo:c6 rdfs:label "Certainty 6"; rdfs:comment "the ancient element is still in situ, and its dimensions and position can be measured".

tdvo:c5 rdfs:label "Certainty 5"; rdfs:comment "the ancient element is not in situ but it has been visually documented in the past and the documentation is still available". tdvo:c4 rdfs:label "Certainty 4"; rdfs:comment "the ancient element is not in situ but it can be geometrically derived from the surviving elements".

tdvo:c3 rdfs:label "Certainty 3"; rdfs:comment "the ancient element is not in situ but it can be visualised according to well accepted standards and precedents".

tdvo:c2 rdfs:label "Certainty 2"; rdfs:comment "the element is not in situ but it can be visualised according to the modeller's experience, knowledge, intuition".

tdvo:c1 rdfs:label "Certainty 1"; rdfs:comment "the element is not in situ and it has been added for communicative purposes".

tdvo:c0 rdfs:label "Certainty 0"; rdfs:comment "the element has not been created for scholarly purpose and does not aim to historical accuracy. However, some characteristics of an original referent can still be recognised".

Represent the relationships between the 3D digital visualisation, its sources, referents and interpretations.

For example: tdvo:isBasedOn rdfs:label "is based on"; rdfs:comment "the shape, dimensions or decoration of the element is based on visual or written information contained in a relevant document describing established practices, standards and rules".

The object of the predicate could be traditional bibliographical references and/or the digital URI of the source and/or the URL of a digital edition of the source such as the ones available on open digital archives (obj:001 tdvo:isBasedOn dbpedia:De_architectura).

tdvo:hasEvidenceIn rdfs:label "has evidence in"; rdfs:comment "the 3D element can be compared with specific verbal or visual evidence such as video/photographic documentation or official excavation records".

The object of this predicate would be archive numbers or bibliographical references identifying physical documents or artefacts, and/or URIs of digital reproductions of them, available on digital databases such as Arachne⁷ or CLAROS⁸. For example, if obj 002 was an element of the 3D visualisation of the Basilica in Pompeii:

obj:002 tdvo:hasEvidenceIn <http://arachne.uni-koeln.de/item/marbilder/2015507>.

tdvo:isMentionedIn rdfs:label "is mentioned in"; rdfs:comment "the visualised building

(or part of it) is mentioned in a ancient (or modern) text".

tdvo:isDescribedIn rdfs:label "is described in"; rdfs:comment "the visualised building (or part of it) is described in a ancient (or modern) text".

The latter predicates could link to bibliographical references and/or to the digital version of ancient texts such as the ones available through Perseus Library⁹.

The main goal of this proposal is not to present a detailed ontology, but to show the potential of the application of Open Linked Data to 3D visualisation, and how such an interaction will change the way 3D visualisation is applied in the study and understanding of cultural heritage. The ontology itself

is not meant to be the work of a single researcher, but the collaborative effort of the different communities of practitioners.

To summarise, the suggested ontology will:

- constrain and standardise the documentation, making it synthetic instead of verbose;
- speed up the recording process thus reducing time/cost and making the documentation more likely to be retained in projects' budgets;
- allow 3D visualisations to join and enrich the growing network of linked digital resources to study the past;
- make 3D visualisations human- and machine-searchable, connecting them with the literary and historical sources that mention the visualised artefact or building;
- allow and encourage comparison of different visualisations and interpretations of cultural heritage, as the same resource (historical, archaeological, literary) will be connected to all the related visualisations that share the same vocabulary;
- allow citations, re-use and peer-review of 3D visualisations, as every 3D element (and its author) will always be identifiable and linkable through the URI;
- contribute to transform 3D visualisation from a univocal display of traditional research to a collaborative virtual environment where different scholars work together not only to implement the content but also to refine the ontology itself.

References

1. www.londoncharter.org
 2. Thanks, for example to APIs such as OpenGL and WebGL
 3. Cf., for example, CIDOC CRM or FRBRoo.
 4. There are a few examples of formal vocabularies dedicated to art and architecture that can be used to describe the represented object. The Thesaurus developed by the Getty foundation, for example, has recently been released in Open Linked Data format (February 2014).
 5. An interesting and useful precedent in using RDF to describe aggregations of files can be found in the open archive initiative.
 6. pleiades.stoa.org
 7. arachne.uni-koeln.de/drupal
 8. www.clarosnet.org/XDB/ASP/clarosHome
 9. www.perseus.tufts.edu/hopper
- Baker, D.** (2012). *Defining Paradata in Heritage Visualization*. In Bentkowska-Kafel,
- A., Denard, H. and Baker, D.** (eds) (2012). *Paradata and Transparency in Virtual Heritage*. Farnham: Ashgate
- Champion, E. and Dave, B.** (2007) *Dialing Up the Past*. In Cameron, F. and Kenderdine, S. (eds) (2007). *Theorizing Digital Cultural Heritage. A critical discourse*. Cambridge, MA: MIT Press
- Favro, D.** (2006) *In the eyes of the beholder: Virtual Reality Recreations and academia*. Journal of Roman Archaeology Supplementary Series Number 61(2006): 321-334
- Forte, M.** (2012) *Cyberarchaeology: a Post-Virtual Perspective*. www.academia.edu/3573281/Cyberarchaeology_a_Post-Virtual_Perspective (accessed 6th March 2014)
- Hermon, S.** (2008) *Reasoning in 3D: a Critical appraisal of the role of 3D modelling and virtual reconstructions in archaeology*. In Fisher, B., and Dakouri-Hild, A. (eds) *Beyond Illustration: 2D and 3D Digital Technologies as Tools for Discovery in Archaeology*. British Archaeological Reports International Series.
- Online Resources:**
- The CIDOC Conceptual Reference Model: www.cidoc-crm.org (accessed 6th March 2014)
- Getty Vocabularies as Linked Open Data: www.getty.edu/research/tools/vocabularies/ld/index.html (accessed 6th March 2014)
- The London Charter: www.londoncharter.org (accessed 6th March 2014)
- Open Archives Initiative: www.openarchives.org (accessed 6th March 2014)
- Pleiades Gazetteer: pleiades.stoa.org (accessed 6th March 2014)

Linked Open Data Technologies and Emblematica Online II

Wade, Mara R

Society for Emblem Studies/University of Illinois, United States of America; mwade@illinois.edu
University of Illinois at Urbana-Champaign, USA

Cole, Timothy

University of Illinois at Urbana-Champaign, USA

Han, Myung-Ja

University of Illinois at Urbana-Champaign, USA

In the context of *Emblematica Online I*, with support from the National Endowment for the Humanities (NEH) and the Deutsche Forschungsgemeinschaft (DFG) from 2009, the University of Illinois and the Herzog August Bibliothek, Wolfenbüttel (HAB), digitized 728 Renaissance emblem books, thereby substantially expanding the digitized corpus. We have digitized approximately 70,000 individual emblems, creating detailed emblem-level metadata for more than 17,000 of these. Each emblem is identified with a globally unique URI (Uniform Resource Identifier) maintained in a shared emblem registry. The *OpenEmblem Portal* prototype was collaboratively designed to provide access to these materials and to demonstrate the feasibility of international repository interoperability.

A new NEH grant from the Historical Collections and Reference Resources program in 2013 focuses on proving the viability of the *OpenEmblem Portal* prototype. This grant features an expansion of the Portal with content from the collections of at least 4 additional institutions. The new grant also supports experimentation with Linked Open Data (LOD) services and RDF-based annotation tools, demonstrating how Semantic Web technologies can facilitate discovery, new modes of scholarly communication, and the more effective use of digitized emblem resources in scholarly research and pedagogy.

To further build the emblem corpus for the *OpenEmblem Portal*, researchers at Illinois will expand the virtual emblem collections in three important ways at varying levels of granularity. We will

- 1) integrate existing digital facsimiles of early modern emblem books from the Getty Research Library, Duke University Library, and the Hathi Trust, in order to demonstrate proof of concept. This increases the number of individual emblem books in the *OpenEmblem Portal* by more than 300 volumes.
- 2) incorporate metadata for 8,231 individual emblems from completed digitization projects at Utrecht University and Glasgow University.
- 3) digitize and create metadata for approximately an additional 100 rare emblem books from the University of Illinois' collections and create emblem-level metadata for them.

By adding the combined number of 8,231 already indexed emblems from Glasgow and Utrecht to 8,000 additional indexed emblems from the University of Illinois, *OpenEmblem Portal* will nearly double the corpus of searchable emblems, increasing the number of indexed emblems currently available from 18,889 to 35,121. *Emblematica Online II* thereby creates an extraordinarily rich resource for all areas of Renaissance Studies.

In comparison to the path-breaking publication of Arthur Henkel and Albrecht Schöne's handbook *Emblematika* (1967), which partially indexed 45 emblem books, *Emblematica Online I* made 723 complete digital facsimiles with 18,889 individual emblems available on the Web. At its conclusion *Emblematica Online II* will expand this corpus to over 1,000 key Renaissance emblem books and present more than 30,000 fully searchable emblems in high quality images indexed by Arkyves according to Iconclass notations and labels for scholars worldwide. The mottos for these emblems will also be searchable in a database of emblem mottos. Additional emblems will be registered and available for browsing, albeit initially without transcription and

Iconclass indexing (pending subsequent projects and possible emblem scholar crowd-sourcing). By analogy to Henkel and Schöne, heralded as a major research accomplishment at the time of its publication, *Emblematica Online II* will accomplish and initiate comparable research activity by a factor of more than twentyfold.

Moreover, the design of the *OpenEmblem Portal* and its use of LOD directly addresses the acute scholarly need to move beyond a catalogue of a large number of digitally available books and will create a significant corpus of richly indexed materials at the sub-book level. While the *OpenEmblem Portal* focuses on emblems digitized from print, emblems also permeate the fine and applied arts. Through LOD best practices and emerging annotation standards, the *OpenEmblem Portal* eventually will allow scholars to link an emblem found in a Bavarian church or a Swedish manor house to a printed emblem. This more granular and semantic approach to describing digitized emblem resources also opens up opportunities to engage graduate and even undergraduate students in the project. For example, undergraduate students at Illinois are meaningfully supporting the *OpenEmblem Portal* by participating in emblem motto transcription and registration.

This poster presents our progress with *Portal* evolution, LOD technologies, and next steps to leverage LOD to facilitate emblem research and pedagogy.

Visualization, Interactivity and Contextualization as Digital Cultural Empowerment: Ancient Egyptian Architectural Terminology OnlinePaper created on 2013-11-01, 08:52

Wendrich, Willeke
wendrich@humnet.ucla.edu
 UCLA

Working in an international setting, such as the international archaeological community in Egypt, brings to the fore that the communication is by definition multi-lingual and, secondly, that international understanding is faced with serious problems of translation. Even if English is used as the modern lingua franca, the different associations connected to terms, based on the various meanings of the terminology in the mother tongue, can cause serious misunderstandings. This is particularly true for communications in which detailed and specialized terminologies are involved. Even in languages that are reasonably close, such as English, French and German, similar sounding terms have developed into a different range of meaning, thus adding to the confusion: a non-native speaker assumes to know the meaning of a term which has the same root, while in reality that meaning does only partly overlap, or diverges completely. The situation becomes even more complex when languages from quite different families, such as Arabic, are included. A translation conveys only part of the term, without touching upon all the linked associations and unspoken meanings. The Arabic term "athar", for instance, is used as a translation for "antiquities" or "monuments", while the core meaning is "remnants" and thus does not convey the strong associations with Greek and Latin antiquity and a world of scholarly and cultural tradition going back to the renaissance that the English term carries with it. Added to that are the very real sensitivities of using the language of former colonizing powers to describe a country's cultural heritage.

The multi-lingual scholarly environment of Egyptian archaeology requires a broad knowledge of specialized terminology, much of which is related to ancient Egyptian architecture. Architectural terms are often used indiscriminately and confusingly. English, French and German terminology

leans heavily on jargon developed by the long tradition of the study of Greek and Roman buildings, such as temples, civic establishments, theatres and other monumental structures. Arabic communications on architecture either use loan words from English, French or Italian, or adopt terms from Islamic architecture, dating to much later periods and a different cultural setting.

Digital Humanities enable an approach that provides possibilities of conveying information in alternative ways. Architectural terminology can and should be understood in its physical, temporal and cultural context. By linking terms to a range of specific buildings they are embedded in place and time. Furthermore, the use of images, colors and interactive interfaces, enables visual learning and avoids lengthy definitions that often obfuscate, rather than clarify. Each multi-lingual equation of terms merits an in-depth discussion and could be the subject of a little conference, but that defies the practical purpose of providing a workable resource to enhance communication.

A cooperation between the German Archaeological Institute Cairo and the University of California, Los Angeles, takes a practical, contextual, visual approach to clarifying architectural terms. Although the function of terminology is partly to enable a certain degree of generalization, studying architecture in its cultural context enables research that teases out temporal and regional differences and developments as well. Furthermore, by including a wide range of building types, including pyramids and workmen's huts, the entire range of building types common in a particular era and region are taken into account.

The multi-lingual terminology uses photographs and drawings of actual building parts, rather than stylized reconstruction drawings to enable this contextual approach. The drawings and photographs of building parts and architectural details are linked to georeferenced plans of specific ancient buildings. Thus architectural terms in four languages (English, Arabic, German, and French) are illustrated within their architectural and cultural context.

The viewer developed to present the terminology is GIS-based and provides not only the building context, but also the geographical context for the building parts under discussion. The building plans are presented in three different forms: actual state, reconstruction and a plan indicating different building phases. In addition elevations, cross sections and architectural details are added, based on existing publications and ground checking in the field. A team of Egyptian architects create AutoCAD drawings of the selected buildings, and is involved in field checking and photography. An interactive interface enables the user to virtually "step into the plan" by knowing the exact location location, direction, angle and tilt of photographs incorporated in the system. This will enable the user to orient him/herself in the building. It is not the same as being physically present in the actual building, but it avoids the complete divorce of the visual realm from the bodily context.

The result is a freely available web resource which empowers users from various cultural backgrounds and different disciplinary training. The Ancient Egyptian Architecture Online (Aegaron) project provides users with vetted architectural drawings and a very practical and accessible way of comparing architectural, archaeological, historical and egyptological concepts.

References

dai.aegaron.ucla.edu

HistoGlobe - Visualising History

Westermeier , Carola
Carola.Westermeier@histoglobe.com
 HistoGlobe, Germany

Our software HISTOGLOBE visualises historical events and developments on a three-dimensional globe and map. With the help of a timeline users can explore history interactively.

In order to emphasize certain aspects of history or in order to make the map more or less detailed several filters can be used. Zooming in and out enables the user to see more or less details. It is designed to be highly user-friendly with an intuitive handling.

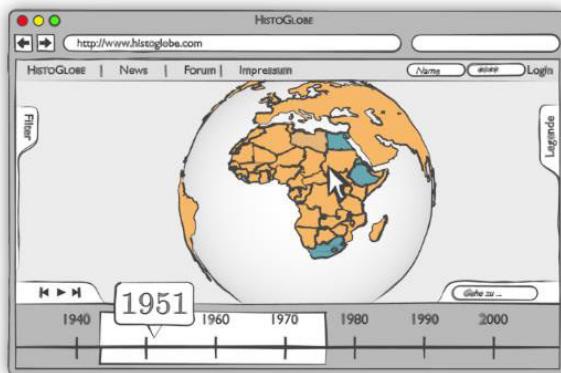


Fig. 1: Simplified HistoGlobe Mockup

HISTOGLOBE is based on a simple web-browser. Thereby, it can be also be used by smartphone and tablet.

The software enables researchers to visualize their topics concerning long and short time developments with a broad regional expansion. Of course, the historical events can be connected with all kinds of (digital) sources, such as documents, pictures and videos. Thereby, HISTOGLOBE enables scientists as well as for example museums and galleries to present their topics in a unique way.

Our current prototype shows the development of the European Union on the basis of nearly 60 events. In order to use the advantages of a visualization tool, countries as a whole can also be highlighted. Thereby, the growth of the European Union is depicted vividly. Picture 2 shows an exemplary date during which the Federal Republic of Germany, Belgium, the Netherlands, Luxemburg, France and Italy are highlighted to show their connection. The stars highlight several events related to the development of the European Union. These events imply a short description and can be matched with sources, such as pictures, videos, audio-files and documents. Of course, the description can also lead to related websites and further information.

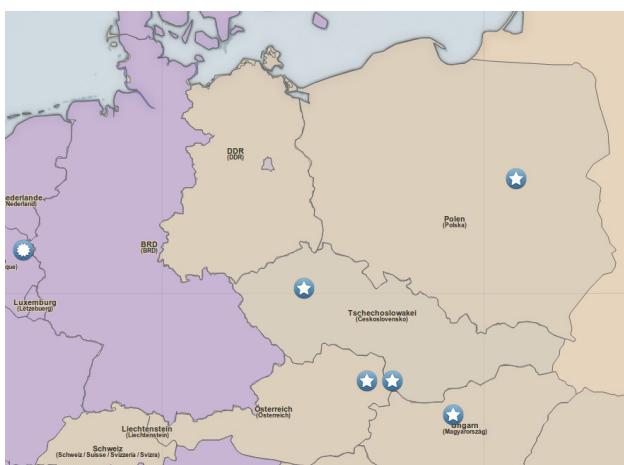


Fig. 2: Developments of countries, taken from www.histoglobe.com



Fig. 3: Single events are connected with multimedia elements

Within the next months we will be able to show different types of events, several filters, connections between events and animations. Of course, HISTOGLOBE can apply different languages.

During the conference we would like to present our visualisation of World War I, which we are currently developing with students at the Justus-Liebig-University Giessen. Conference visitors will be able to use the visualisation and navigate through the course of history using a laptop. If visitors are interested, we are open to show how historical events are implemented in the software and which techniques were used.

In addition, we would like to show how HISTOGLOBE has been in use in several museums in Germany and how it was already used by teachers and students to create their own version of HISTOGLOBE. Based on these examples, we would like to invite academics to discuss how future research could be visualised using HISTOGLOBE.

Our multidisciplinary team consists of six (PhD-) students. En detail, we are three computer scientists, two historians and one business student. Our project is supported by the Heinz Nixdorf Foundation and the Foundation of German Business. The first prototype was developed at the Abo Akademi in Turku, Finland. We already cooperate with the Bauhaus- University Weimar and the Justus-Liebig-University Giessen.

Semantic Blumenbach: Exploration of text-object relationship with semantic web technologies in the history of science

Wettlaufer, Jörg

jwettla@gwdg.de

Göttingen Centre for Digital Humanities (GCDH), Akademie der Wissenschaften zu Göttingen (ADWG)

Blumenbach-online, a project of the Göttingen Academy of Sciences and Humanities, started in January 2010 and aims at both digitizing and presenting the writings and collections of the influential Gottingen physician and naturalist Johann Friedrich Blumenbach (1752-1840), one of the founding fathers of physical anthropology, online. To date, almost half of the textual material (77.000 pages altogether) and roughly a quarter of the collections have been digitized and converted into TEI-encoded texts or entered into a database. It is through an exploration and application of Semantic Web technologies in a spin-off project called "Semantic Blumenbach" that we hope to establish robust and powerful methods for presenting and providing heterogeneous machine-readable linked data for Blumenbach-online.

Two major tasks have been completed so far. The first is carrying out Named Entity Recognition (NER) on the TEI P5 Tite¹ encoded full-texts that have been provided to Semantic

Blumenbach² by Blumenbach-online. These texts lacked the semantic markup e.g. for places, persons and objects from the natural history domain. In addition, we had to deal with historical and irregular orthography of multilingual texts from the second half of the 18th century. Currently we are able to recognize precisely (96%) most (96%) of the technical terms that appear in the text using a list-based algorithm. This algorithm is also able to detect binomial entities from the Linnaean taxonomy, even when they appear as separate strings in different parts of the text. For modeling the relationship between entities in the text and metadata in the collection, we use the WissKI Framework for scientific communication (www.wiss-ki.eu) that allows presenting and using data from various sources in a robust and open system, which is both scalable and reusable by other projects. With the help of the Erlangen CRM Ontology³, an OWL-DL 1.0 implementation of the CIDOC CRM⁴ and a special application ontology, we model the semantic relationships between objects described in TEI-encoded texts and metadata of these objects.⁵ We particularly focus on place names, persons and special terms from the natural history domain, including the Latin names of animals and geological objects and construct the relationship between both types of data by using our NER to encode reference strings in the TEI text.

The Erlangen CRM provides a way to classify these objects in a meaningful way and to model the relationship between the occurrence of the objects in the writings of Blumenbach and the University of Göttingen's collections. With the help of colleagues from the WissKI Project at Erlangen and Nurnberg we have been able to develop new modules for the Drupal-based system to ingest the TEI and triplify the metadata that we created in the texts. Following a policy of Open Access and Linked Open Data, we will test and implement ways to generate and publish results of academic research in a way that it can be reused in other contexts and by other researchers. Finally, we plan to use a full-text search index (Apache solr) to make both texts and object-related data available in a way that allows both triplyfied metadata and XML full-text to be searched efficiently.

URL: dhfv-ent2.gcdh.de/blumenbach/wisski
Username and password available on request.

References

1. www.tei-c.org/release/doc/tei-p5-exemplars/html/tei_tite.doc.html
2. Wetzlaufer, Jörg & Thotempudi, Sree Ganesh (2013): *Poster - NER in historical Text corpora*. Lessons learned so far. 4.-6.03.2013, Mehr Personen – Mehr Daten – Mehr Repositorien, Tagung des Personendatenrepositoriums der BBAW, Berlin. www.gcdh.de/index.php/download_file/view/168/405
3. erlangen-crm.org
4. "CIDOC CRM," n.d. www.cidoc-crm.org/index.html.
5. C.f. www.tei-c.org/SIG/Ontologies/meetings/m20131003.html

"Tout ce qui n'est point vers, est prose" : Raymond Queneau's Matrix Analysis of Language, Syntactic Stylometry, and Exploratory Programming

Wolff, Mark
wolffm0@hartwick.edu
Hartwick College

Introduction

Established techniques in stylometry typically measure word and ngram frequencies with limited consideration of syntax. While it is often easier to access and interpret statistically significant words in a text, an analysis of syntax alone can provide interesting and unexpected results. The analysis presented here represents what Nick Montfort calls exploratory programming, where "there's no specification or problem to be solved, but there are things to be discovered."¹ An initial research question can be a pretext for exploring computation as a means of discovery rather than modeling.

Raymond Queneau's Matrix Analysis of Langauge

Raymond Queneau, a founding member of the Oulipo who recognized the potential of computation for literary analysis and creation, developed a technique for measuring a text's syntax. In a paper he published in 1964², Queneau explored the mathematical properties of a system of tagging parts of speech according to 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 word group such as a sentence, one can construct two matrices where the first matrix contains all formatives and the second all signifiers. If a word group contains two consecutive formatives or signifiers, one can use a unitary element in order to construct the matrices:

$$\begin{array}{c} \parallel Le \quad 1 \quad a \quad bien \quad la \quad 1 \parallel \times \\ \parallel vilain \quad chat \quad 1 \quad mangé \quad belle \quad souris \parallel \end{array} = \begin{array}{l} (Le \times vilain) + (1 \times chat) + (a \times 1) + \\ (bien \times mangé) + (la \times belle) + (1 \times souris) \end{array}$$

Fig. 1: The sentence "Le vilain chat a bien mangé la belle souris" can be represented as the product of two matrices.

The product of a formative and a signifier is a bi-word. By adopting the conventions that neither (1 x 1) nor (A x 1) + (1 x B) are allowed, one avoids uninteresting or redundant bi-words. Any sentence can therefore be transformed into a sequence of pairs of words, and each pair is either a bi-word (B), a formative (F), or a signifier (S). According to this schema, the sentence in Fig. 1 can be rendered as

B S F B B S.

Syntax and Textual Signals

In previous research I explored Queneau's matrix analysis as a by-product of a more fundamental approach to ludic experimentation in computational literary analysis.³ Initial results showed that matrix analysis could be used to attribute authorship with reasonable success. With the development of the stylo package of stylometry tools for the R programming language,⁴ I combined matrix analysis with cluster analysis in order to determine if an authorial signal could be detected through syntax alone. Using a corpus of 17th-century French plays compiled by Fièvre⁵ and preprocessed by Schöch,⁶ I transformed the plays into sequences of the letters F, S, B and P using Schmid's TreeTagger parser.⁷ I added the letter P to Queneau's schema to designate punctuation, which interrupts the flow of words and allows for occurrences of F P S instead of F S (which would be a bi-word, or B). Spaces were inserted between the letters to facilitate word token analysis (instead of character analysis). The first lines of Molière's *Les Femmes savantes*⁸

*Quoi ? le beau nom de fille est un titre, ma soeur,
Dont vous voulez quitter la charmante douceur,
Et de vous marier vous osez faire fête ?
Ce vulgaire dessein vous peut monter en tête ?*

are thus rendered as the following sequence of letters:

F P B S B F B P B P S B S B S P S F B B S S P B S B S B P

A cluster analysis of the corpus reveals that that the dominant signal is not authorial but formal, depending on whether a text is written in verse or prose:

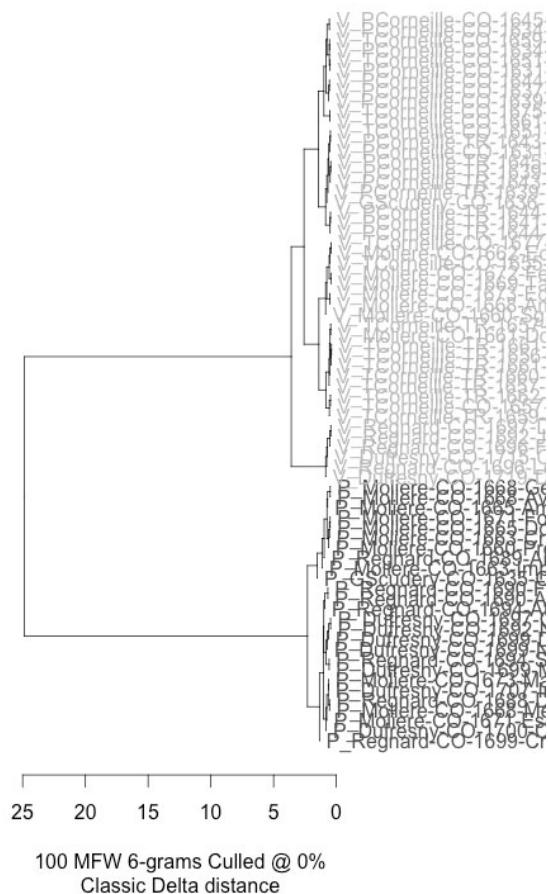


Fig. 2: Cluster analysis of 17th-century French theatre texts reduced to their syntactic structure according to Queneau's schema for matrix analysis.

The corpus clustered perfectly into groups of verse texts (light) and prose texts (dark). These results are surprising, given that traditional verse is determined by meter and rhyme and not explicitly by syntax. Because texts in verse follow the convention of capitalizing the initial letter of the first word of each line, the TreeTagger parser occasionally identified conjunctions and other formatives as proper nouns, tagging them erroneously as signifiers. To see if capitalization affected significantly the clustering, I lowercased every letter (thus masking all proper nouns):

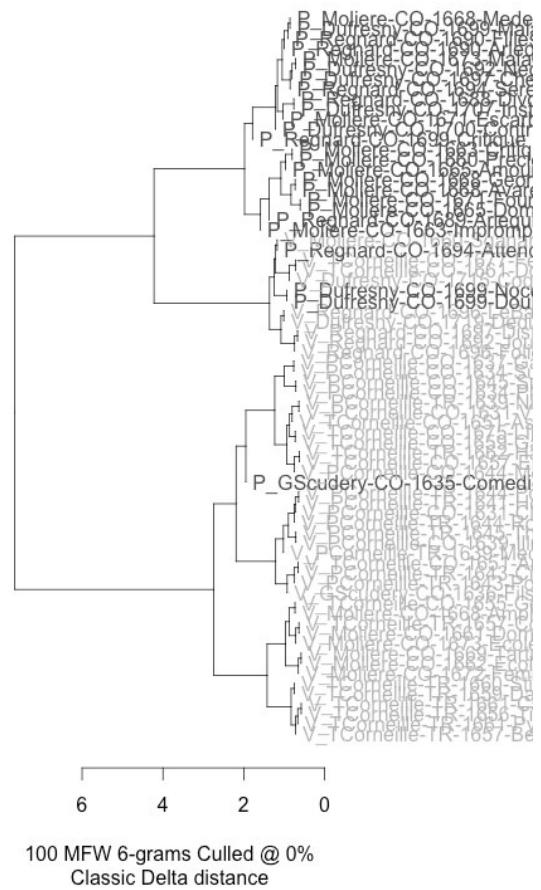


Fig. 3: Cluster analysis of 17th-century French theatre corpus with lowercased texts.

The results are nearly the same. A more accurate analysis would require documents encoded to identify proper nouns.

Principal Component Analysis of Syntax Sequences

In order to examine more closely the syntactic structures differentiating verse from prose, I adopted a technique developed by Khmelev and Tweedie using Markov chains of letters to detect low-level sequence patterns.⁹ Given any text, one can produce a transition matrix that represents the frequencies of Markov chains of bigrams based on Queneau's schema. Here is the transition matrix for *Les Femmes savantes*:

	S	F	B	P
S	0.2183949	0.2482244	0.2769886	0.2563920
F	0.0000000	0.3890392	0.4995489	0.1114118
B	0.2373926	0.2155913	0.2045805	0.3424356
P	0.4042477	0.3707703	0.2221022	0.0028798

This produces sixteen possible bigram combinations, although in reality there are only fifteen because FS never occurs (FS = B). We can consider the frequency of each bigram as a distinct measurement of a text and then analyze all the texts in the corpus as 15-dimensional vectors (this approach is similar to that of Hirst and Feiguina¹⁰). I reduced the vector space to three principal components and generated the following three-dimensional triplot (projected here as three biplots):

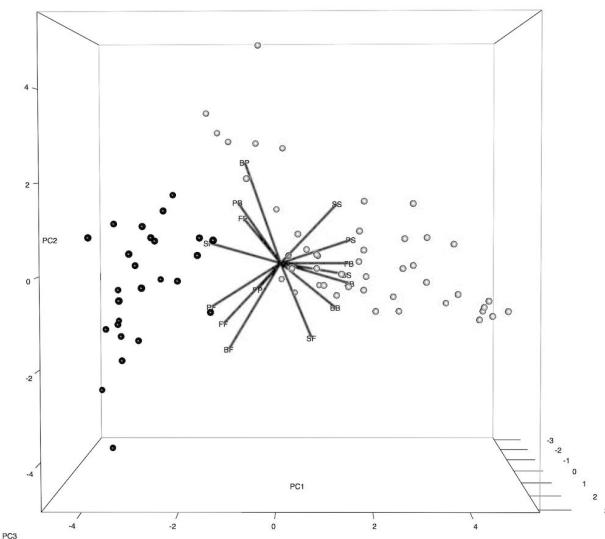


Fig. 4: Projection of PC1 and PC2 from a PCA triplot.

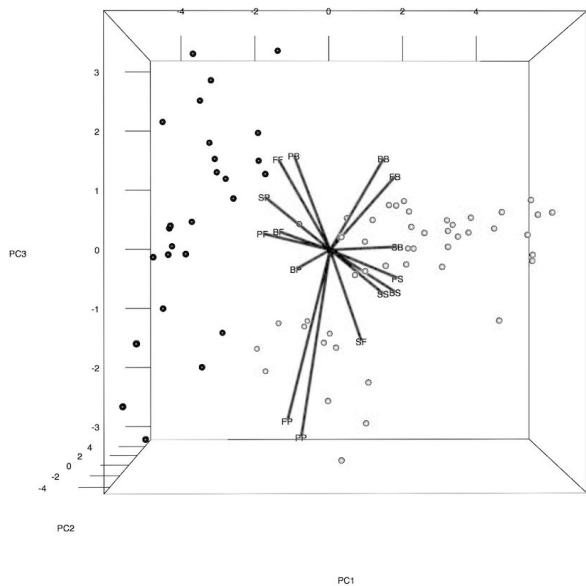


Fig. 5: Projection of PC1 and PC3 from a PCA triplot.

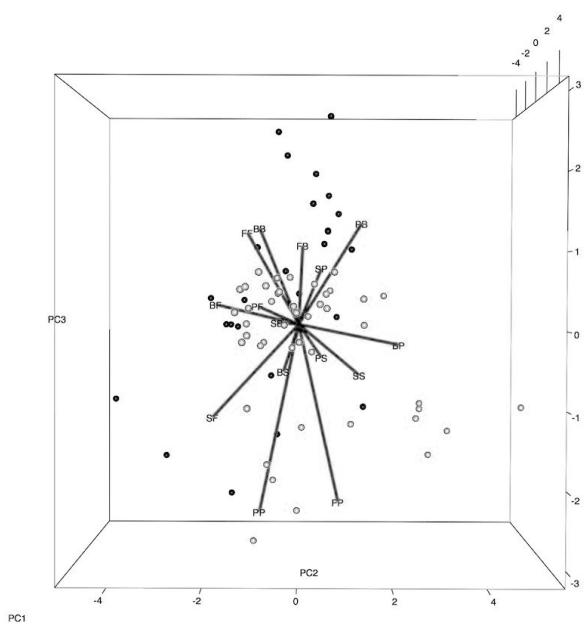


Fig. 6: Projection of PC2 and PC3 from a PCA triplot.

The significant rotations for PC1 are SP, PF, FF, BF and FP, correlated negatively with BB, SS, BS, FB, SB and PS; those for PC2 are BF, SF and FF, correlated negatively with FP, SS, PB and BP; and for PC3 the significant rotations are PP, FP and SF, correlated negatively with FB, FF, BB and PB. These results are preliminary but Fig. 7 clearly shows how prose and verse texts separate in the triplot:

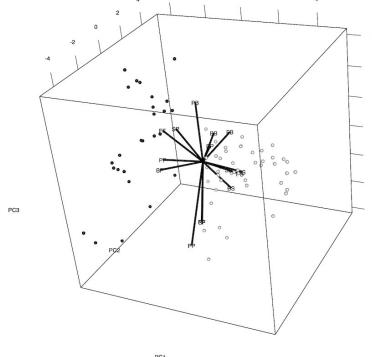


Fig. 7: Angled projection of PCA triplot.

There is a higher tendency among verse texts toward SS (consecutive signifiers), PS (initial signifiers after punctuation), SB and BS (signifiers and a bi-words in either order). Prose texts tend toward lower frequencies of SP (signifiers with no preceding formative), followed by punctuation), FF (consecutive formatives), PB (initial bi-words after punctuation), PF (punctuation followed by formative) and BF (bi-words followed by formative). From these observations we can extrapolate further and say tentatively that in the syntactical structure of a text, verse tends to feature signifiers and prose tends to avoid formatives. These results confirm earlier analyses of classical French plays by Beaudouin and Yvon who detected high frequencies of nouns in the sixth and twelfth metrical positions of alexandrine verse.¹¹

Conclusion

It would seem that the Maître de Philosophie in Molière's *Bourgeois gentilhomme* (II, 4) is not entirely risible in explaining the difference between verse and prose to Monsieur Jourdain.¹² There appears to be a definite measurable difference between these two text forms, at least in French. What is remarkable with this finding is that the difference does not depend on specific word choice, meter or rhyme, even though those are the qualities readers appreciate in verse. I have completed a similar analysis with the ABU corpus¹³ (over 200 works in French spanning many centuries) and the results are comparable:

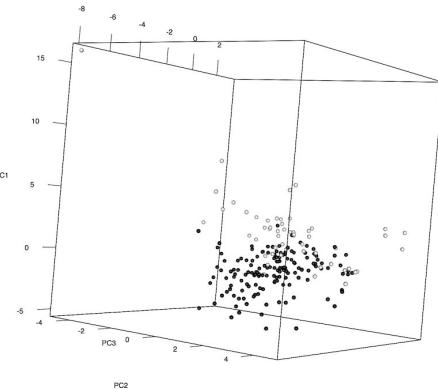


Fig. 8: PCA triplot of the ABU corpus.

Cluster analysis and principal component analysis indicate that verse and prose are measurably different according to a purely syntactical analysis, with no explicit reference to semantics, phonetics or scansion. This discovery resulted not from an initial hypothesis about the relationship between syntax and text type but from exploratory programming, where a statistical technique commonly used to test authorship was applied to a purely syntactical transcription of texts. The investigation of an initial hypothesis (that authorship can be attributed to syntactical patterns) led to an entirely different conclusion through experimentation with computational techniques. One could pursue this research further by using Queneau's matrix analysis of language to test liminal works such as Baudelaire's *Petits poèmes en prose*¹⁴ to determine if an example of modernist poetry classifies as verse or prose. It should not be overlooked, however, that computational text analysis can produce interesting results through serendipity.

References

1. **Montfort, Nick.** (2014) *Exploratory Programming*. Critical Code Studies Working Group. Web. 7 March 2014.
2. **Queneau, Raymond.** (1964) *L'Analyse matricielle du langage*. Etudes de linguistique appliquée 3: 37–50. Print.
3. **Wolff, Mark.** (2007) *Reading Potential: The Oulipo and the Meaning of Algorithms*. 1.1. Digital Humanities Quarterly. Web. 30 Oct. 2013.
4. **Eder, Maciej, Mike Kestemont, and Jan Rybicki.** (2013) *Stylometry with R: a Suite of Tools*. Digital Humanities: Conference Abstracts. Lincoln, NE: 2013. 487–89. PDF file.
5. **Fièvre, P.** (ed. 2007-2013). *Théâtre classique*. Web. 30 Oct. 2013.
6. **Schöch, Christof.** (2013) *Data Is All You Need: Documentation for 'Fine-Tuning Our Stylometric Tools' at DH2013*. The Dragonfly's Gaze. Web. 30 Oct. 2013.
7. **Schmid, Helmut.** *TreeTagger: a language independent part-of-speech tagger*. Institute for Natural Language Processing, University of Stuttgart. Web. 30 Oct. 2013.
8. **Molière.** (1672) *Les Femmes savantes*, comédie. **Fièvre**. Web. 30 Oct. 2013.
9. **Khmelev, Dmitri V., and Fiona J. Tweedie.** (2001) *Using Markov Chains for Identification of Writers*. Literary and Linguistic Computing 16.3: 299–307. Web. 30 Oct. 2013.
10. **Hirst, Graeme, and Ol'ga Feiguina.** (2007) *Bigrams of Syntactic Labels for Authorship Discrimination of Short Texts*. Literary and Linguistic Computing 22.4: 405–417. Web. 30 Oct. 2013.
11. **Beaudouin, Valérie, and François Yvon.** (1996) *The Metrometer : a Tool for Analysing French Verse*. Literary and Linguistic Computing 11.1: 23-31. PDF file.
12. **Molière.** (1670) *Le Bourgeois gentilhomme, comédie-ballet*. Fièvre. Web. 30 Oct. 2013.
13. **ABU (Association des Bibliophiles Universels).** (1993-2013) *La Bibliothèque Universelle*. Web. 30 Oct. 2013.
14. **Baudelaire, Charles.** (1869) *Le Spleen de Paris ou Petits Poèmes en Prose*. Litteratura.com. Web. 30 Oct. 2013.

A Text Encoding Support System for Pre-modern Japanese Historical Materials

Yamada, Taizo

t_yamada@hi.u-tokyo.ac.jp

Historiographical Institute, The University of Tokyo

Inoue, Satoshi

inoue@hi.u-tokyo.ac.jp

Historiographical Institute, The University of Tokyo

1. Introduction

Reading comprehension of historical materials is one of important elements in historical study. The results of the reading comprehension should be encoded as texts; however, in Japanese historical study amount of the texts is a few rather than of digital images. Almost encoded texts are not shared and there are no rules for text structuring.

In the study due to structuring encoded texts automatically and sharing the texts among researchers, we developed a text encoding support system for pre-modern Japanese historical materials, especially Japanese medieval period. The features of our system are follows: web-based system, automatic text structuring, text editing, text sharing and support for reading the characters in the materials. Our system doesn't have and manage any material's catalogues. We suppose that the system uses a ready-made system to search catalogue. Particularly, we use "Catalogue database of holding materials"¹⁵ (called "HICAT") in Historiographical Institute the University of Tokyo.

2. Our System

2.1. Basic Methods

Our system has 2 methods; search method and authoring method. The search method allows a user to search for images, texts and annotations. An annotation assignment is one of important work in encoding for research of Japanese history. Our system can deal with following 2 annotation types: marginal note and format note. In the study we defined that a marginal note is a description of "a result of reading comprehension or research" and disappear in the material. Examples of marginal note are personal name, location name, correction and so on. Format note indicates descriptive pattern for strings (e.g. erasure, divide note,...) or lines (address, title, subject,...).

Editing the texts and the annotations can be supported by the authoring method. Using the authoring method, the system starts text structuring automatically as soon as a user edit a text. If the text editing is finished and the text is committed, then new version of the text is created. A version is identified by a user ID, modified time and image ID (as URI). A user can use the previous version and the versions of other users. If the user edits other user's version, the new version will be created. The new version takes over all annotations in original version and can be edited freely. Therefore, the method of text reuse never violates other user's text.

2.2. Attempt of converting into TEI

The system can output XML document as the result of text encoding. The structure is useful only in our system, because the structure is specialized in the system. We think an encoded text should be outputted in a general format when the text is used outside our system. Because TEI P5² is "de facto standard" of text encoding in Humanities, we attempted convert our text into TEI P5. We carefully treat the expression of the line and the annotation in the conversion, because in our system text is represented as a set of lines and annotations. For the expression of the marginal notes as personal name, place name, and correction, we use <persName>, <placeName>, and <choice> respectively.

Moreover, we consider automatical assignment of opener and closer in the text. We analyze a form pattern of Japanese historical materials, and the assignment is realized by the basis of the results.

2.3. Reading Support Method

Since the encoding a historical material is very hard, the researcher of Japanese history is needed training or practice for a long time. In order to support the encoding, we provide a suggestion method for support of inputting character. When a user input string in a text field on our system, the suggestion

method presents a candidate character which appears after current inputted string. The method is realized by character n-gram model. A learning data of the n-gram is constructed by texts extracted from fulltext database of Historiographical Institute. In order to improve the precision of the suggestion method, we use Modified Kneser-Ney Smoothing method³. We experiment for the confirming the performance of the suggestion method. As the experimental result, the hit ratio whether a set of candidate character in top 20 includes a correct is 0.72. The ratio might seem to low, but it can be effectively used in the actual work.

3. Conclusion

Our system has been developed for managing texts which are represented as results of reading comprehension. We believe that the most important element of the study is to provide an environment in which researchers of Japanese history can encode texts pleasantly and comfortably. In order to achieve it, we'd like to improve the expressiveness of texts and performance of methods in the system.

References

1. *Databases of Hl.* wwwap.hi.u-tokyo.ac.jp/ships/.
2. *TEI guidelines*. www.tei-c.org/Guidelines/P5/.
3. F. James (2000), *Modified Kneser-Ney Smoothing of n-gram Models*, Technical report, RIACS Technical Report 00.07, www.riacs.edu/navroot/Research/TRpdf/.

Taco: A Metadata System for Hierarchical Structured Data Collections

Zastrow, Thomas

thomas.zastrow@rzg.mpg.de
RZG

Gross, Karin

karin.gross@rzg.mpg.de
RZG

1. Introduction

Today, any modern file system offers the possibility to create hierarchical nested folder structures with an arbitrary depth. This leads often to large accumulations of data, where the only regulative element is the hierarchy of folders. Such a directory structure represents meta information about the data it contains, but because this information is not bound to specific datastreams of bits, it is often not included into traditional metadata formats which are used to describe data stored in files¹. Taco (short for „Tags & Components“) is a metadata system which allows to assign metadata directly to the folders of a file system.

1.1. Overview

Tags are community-specific, predefined attribute-value pairs of a specific data type, while a component is a collection of tags, contextually affiliated with each other. Such a component can then be assigned to a folder, describing the content of the folder as a whole. It doesn't matter how deep or where exactly in the file system hierarchy the folder is located. The metadata is stored as key-value pairs in plain text files. These files can be converted to any other metadata format or exported via OAI-PMH². After the creation of the Taco components, a harvester parses the whole hierarchy recursively and indexes all assigned

metadata components. The assignment of components to folders is optional, with one exception: At the root level of the projects file system hierarchy, a „header“ component is mandatory. It represents the top-most entry point for a specific project. It contains general information like owner, contact information and a description about the whole project. It is used as an anchor for all other components: all non-header components can be seen in relation to the header.

The Taco System does not prohibit the use of common metadata formats like CMDI for describing individual file(s) – optionally, these file-based metadata formats can be included into the Taco system.

The Taco system consists of three parts (Figure 1): (I) TacoEdit, a desktop application for assigning components to folders, (II) TacoHarvest, a harvesting application which collects the components inside a folder hierarchy and stores them in a (relational) database and (III) TacoBrowse, a web application that provides simple access to that database. All three parts are built around predefined and community-specific tags and components.

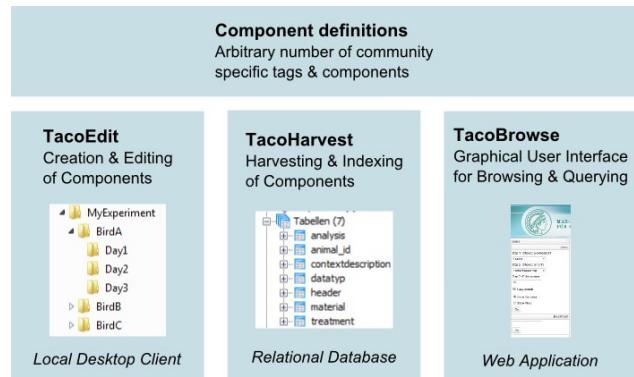


Fig. 1: The three parts of the Taco System

2. The Taco System

2.1. The Editor

TacoEdit is a user-friendly desktop application. It loads the predefined tags and components and offers an overview of the project's data set, starting with the current project's root folder. From here, the user can create new and edit existing components underneath the root directory.

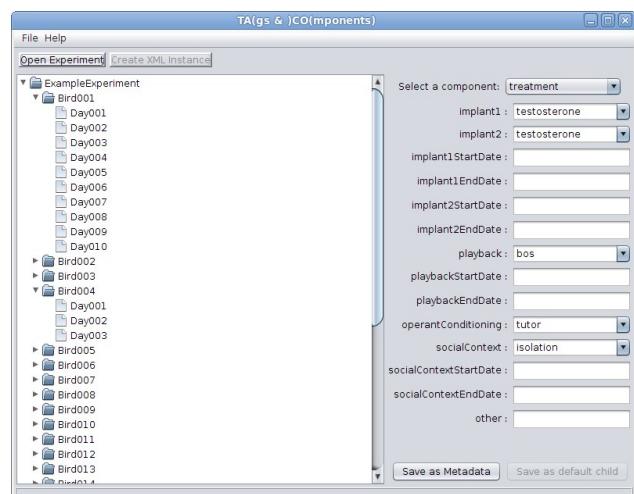


Fig. 2: The TaCo Editor Application

2.2. Harvesting & Indexing

After the user has created metadata components, the second part of the tool chain can be applied to the folder hierarchy.

It scans the folder recursively and extracts the metadata, compiling a tree of metadata components which typifies the whole metadata of a project. This tree of metadata can be exported to a number of XML based metadata formats like CMDI or odML⁴. In addition, TacoHarvest can automatically generate a relational database schema from the predefined list of components and insert the specific metadata via SQL commands. This database is the basis for the third part of the tool chain, TacoBrowse.

2.3. Browsing & Exploring

TacoBrowse is a web application. It can be used to access the database which was created by TacoHarvest. It offers convenient „wizards“ to perform semi-automated search requests to the database. The result of such a request is always one or several folders, represented by the assigned metadata components and its content in form of files and other folders. The user has the choice to either display the metadata component or the files, stored in the result folder(s).

3. Use Cases

The Taco System was originally designed for the Max Planck Institute for Ornithology. Here, it will be used to assign metadata to a very large collection of projects and files (ca. 70 millions files), stored in a migrating file system. The Taco System is designed to be independent of a specific research discipline and can be used wherever data is stored in a hierarchical way. This applies not only in natural sciences, but also to many humanity disciplines which are dealing with large collections of texts, images, videos or other digital content.

4. Conclusion

The Taco System is a complete software solution for assigning, editing and querying metadata assigned to folders in a file system hierarchy. With the current implementation, it is possible to search for tags or components in relation to a project's root directory, represented by the metadata header. In upcoming versions, it will be possible to define explicitly relations between the components and include these into the queries.

References

1. Schmidt, Ingrid (2004). *Modellierung von Metadaten*. In: Henning Lobin; Lothar Lemnitzer: Texttechnologie. Perspektiven und Anwendungen. Stauffenburg, Tübingen, ISBN 3-86057-287-3, S. 143–164.
2. Open Archives Initiative Protocol for Metadata Harvesting: www.openarchives.org/pmh/
3. Broeder, D., Schonefeld, O., Trippel, T., Van Uytvanck, D., and Witt, A. (2011). *A pragmatic approach to XML interoperability the component metadata infrastructure (CMDI)*. In Balisage: The Markup Conference 2011, volume 7.
4. Grewe J., Wachtler T. and Benda J. (2011). *A bottom-up approach to data annotation in neurophysiology*. Front. Neuroinform. 5:16. doi: 10.3389/fninf.2011.00016

Author Index

Abi Haidar, Alaa	422
Adelaar, Nadine	439
Aihara, Kenro	475
Aitken, Brian	424
Alex, Beatrice	198
Alexander, Marc	67, 68
Algee-Hewitt, Mark	70, 72
Allemand, Frédéric	77
Almajai, Ibrahim	231
Almogi, Orna	486
Alvarado, Rafael	73
Alves, Daniel	306
Anderson, Jean	68
Anderson, Wendy	67, 424
Andert, Martin	469
Andrew, Liam	12, 73
Andrews, Tara	64
Andriani, Paola	378
Andréys, Clémence	75
Ankenbrand, Rebecca	425
Anthony, Eleanor Chamberlain	76
Antonacopoulos, Apostolos	185
Antonijevic, Smiljana	76
Armand , Cécile	426
Armaselu, Florentina	77, 80
Arnold, Taylor	382
Arrowsmith, Colin	396
Arvanitopoulos Darginis, Nikolaos	83
Ataoguz, Kirsten	427
Atenas, Javiera	517
Auvil, Loretta	44, 130, 185, 508
Bailey, Scott	502
Bailón-Moreno, Rafael	301
Baker, James	15, 427
Baldi, Vania	285
Bannerman, Gavin	211
Baranda Leturio, Nieves	500
Bardiot, Clarisse	58
Barker, Elton	14
Barker, Elton T. E.	355, 401
Barney, Brett	85
Baron, Alistair	68, 179
Barr, Rebecca	546
Barrón, José Francisco	306
Bartalesi, Valentina	378
Bartz, Thomas	428
Bauer, Jean	431
Bauman, Syd	385, 432
Baumgardt, Frederik	432
Bazarnik, Katarzyna	281
Büchler, Marco	220
Beel, David	86
Beer, Nikolaos	42
Beißwenger, Michael	428
Belli, Jill	87
Ben-Shalom, Adiel	89
Benardou, Agiatis	4, 433
Benes, Jakub	434
Benner, Drayton Callen	91
Beretta, Francesco	116, 468
Bergenmar, Jenny	497
Berger, Joachim	114
Bernardini, Caterina	425
Berra, Aurélien	62
Berti, Monica	432
Bevan, Andy	294
Bögel, Thomas	7
Bögel, Tina	117
Bhattacharyya, Sayan	508
Bianchini, Claudia S.	94
Biber, Hanno	50, 436
Binder, Frank	95
Birke, Peter	436
Bischoff, Kerstin	436
Blümm, ,Mirjam	526
Blumtritt, Jonathan	523
Bodard, Gabriel	5, 20, 439
Bode, Katherine	98
Boggs, Jeremy	502
Bohl, Benjamin W.	42
Bolli, Laurent	543
Bonacchi, Chiara	294
Bonch-Osmolovskaya, Anastasia	100
Boot, Peter	102
Booth, Alison	103
Bordalejo, Barbara	54
Borek, Luise	105
Borel, Clément	75
Borg, Trevor	107
Borgia, Fabrizio	94
Bornet, Cyril	222
Borries, Tony	44
Boschetti, Federico	109
Boschetto, Sylvain	468
Bourg, Chris	15
Bouvier, David	445
Boyd, Douglas	522
Bradley, John	20, 111
Brückweh, Kerstin	436
Brennan, Mallory	163
Broadwell, Peter M.	271
Brookes, Stewart	539
Brotnov Eckstrom, Mikal	425
Broun, Dauvit	111
Brown, David	120
Brown, David Michael	38
Brown, Monica	296
Brown, Susan	272, 439, 441
Brown, Travis	499
Bruhn, Kai-Christian	6
Brundin, Michael	441
Brunner, Annelen	112
Bubenkofer, Noah	29
Buedenbender, Stefan	443
Buomprisco, Giancarlo	540
Burch, Thomas	114, 443
Burrows, Toby	115
Burton, Valerie C.	374
Butez, Claire-Charlotte	116, 468
Butt, Miriam	117
Cade-Stewart, Michael	120
Caldas, Natalia	120, 413
Campagnolo, Alberto	458
Capitanu, Boris	130
Carvais, Robert	123
Caton, Paul	125
Cawthorne, Jon	535
Celano, Giuseppe	432
Chambers, Sally	4, 8
Champion, Erik	4
Chao-lin, Liu	404
Chartier, Ryan	272
Chartrand, James	441
Chateau, Emmanuel	123
Chatzidiakou, Nephelie	433
Chen, Chun-Wen	444
Chen, lihua	207
Choueka, Yaacov	89
Christian-Emil, Ore	20
Christian-Lamb, Caitlin	529
Christie, Alex	126
Christy, Matthew	129, 130, 184, 473, 498
Chung, Chia-Hsuan	207
Ciotti, Fabio	492
Ciula, Arianna	10, 257
Clark, Konnor	168
Clement, Tanya	44, 132
Clifford, Jim	198
Clivaz, Claire	v, 31, 335, 445
Coate, Bronwyn	396
Coble, Zach	466

Cocquyt, Tiemen	391
Cole, Timothy	465, 552
Coles, Katharine	446
Compeau, Timothy	227
Constantopoulos, Panos	11
Cornell, Elizabeth	149
Couture, Stéphane	134
Craig, Goodere	308
Craig, Hugh	135
Crane, Gregory R.	432
Creel, James	416
Crombez, Thomas	50
Crompton, Constance	27
Crowther, Charles	46
Croxall, Brian	327
Cruces-Rodríguez, Antonio	301
Cummings, James	10, 448
Czmiel, Alexander	10
Daftary, Darius	163
Dahl, Jacob	46
Dallachy, Fraser	68
Dallas, Costis	4, 11
Daniel, Pitti	20
Das Gupta, Vinayak	137
DataColtrain, James Joel	138
Davidson, Alwyn	396
Davis, King	533
Day, Jia-Ming	444, 449
Day, Shawn	26
De Marsico, Maria	94
De Paepe, Timothy	450
Dean, Will	543
Dee, Stella	432
Deegan, Marilyn	140
Dekker, Ronald	368
Del Grosso, Angelo Mario	109
Del Rio Riande, María Gimena	174
Dellwo, Volker	231
Depauw, Mark	439
Dershowitz, Idan	451
Dershowitz, Nachum	89, 451, 486
Derven, Caleb	141
DiCamillo, Peter	432
Di Pietro, Chiara	501
Dighton, Desiree	143
Dillen, Wout	453
Dillon, Elizabeth Maddock	157
Dimara, Evanthia	434
Dimmit, Laura	144
Dipper, Stefanie	29
Dobson, Teresa	296, 439
Dombrowski, Quinn	105, 453
Donaldson, Chris	179
Downie, J. Stephen	508
Doyle, Benjamin	477
Draxler, Christoph	37
Düring, Marten	407
Drucker, Johanna	64
Duguid, Timothy	498
Dumont, Stefan	454
Dunn, Stuart	22
Dunning, Alastair	4, 146
Durity, Anthony	147
Dutta, Nandita	413
Dye, Dotty J.	149
Dyens, Olliver	455
Earhart, Amy	150, 213, 515
Eberle-Sinatra, Michael	455
Eder, Maciej	135, 281, 456
Edwards, Richard	152
Egan, James	431
Eidem, Laura	70
Eide, Øyvind	152
Eisenstein, Jacob	154
Elliott, Devon	44
Elwert, Frederik	457
Emery, R. Douglas	386, 458
Engel, Maureen	459
Entrup, Bastian	95
Erez, Eden Shalom	411
Ermolaev, Natalia	50
Estill, Laura	460
Fallaw, Colleen	465
Fankhauser, Peter	461
Farquhar, Adam	427
Faulk, Katherine	217
Fechner, Martin	454
Fendt, Kurt	12, 463
Fenlon, Ali	528
Fenlon, Katrina	465
Ferhod, Djamel	468
Ferrer, Carolina	156
Fertmann, Susanne	535
Filarski, Gertjan	156, 368
Fink, Kristina	443
Finlayson, Mark	340
Fisher, Mark	495
Flanders, Julia	157, 528
Folsom, Jamie	12, 463
Foradi, Maryam	433
Forest, Dominic	292
Fornaro, Peter	159, 335
Forstall, Christopher	168, 524
Fourmentraux, Jean-Paul	58
Fournier, Melanie	330
Fragaszy Troyano, Joan	36, 466
Frank, Queens	443
Franzini, Emily	433
Franzini, Greta	5, 433
Fraternali, Piero	407
Friedrich, Vivien	443
Frizzera, Luciano	272, 459
Fujinaga, Ichiro	186
Fuller, Simon	160
Funk, Stefan	544
Furuta, Richard	229
Gabaccia, Donna	346
Galina Russell, Isabel	306, 467
Galloway, Pat	533
Ganahl, Simon	163
Ganascia, Jean-Gabriel	422
Garcia Moron, Javier	407
Gatica-Perez, Daniel	165
Gawley, James	168
Gedzelman, Séverine, Sonia	468
Geßner, Annette	220
Georgieff, Lukas	169
Gervas, Pablo	340
Gießler, André	314, 469
Gil, Alex	15, 33
Gius, Evelyn	7, 22
Gladstone, Clovis	334
Godinez, Marco Antonio	306
Gold, Matthew	453
Gold, Nicolas E.	471
Gold, Valentin	117
Goldman, Jean-Philippe	231
Goldstone, Andrew	171
Gomes, Mariana	472
Gonzalez, Desi	73
González-Blanco, Elena	54, 174, 306
Gooding, Paul	175
Gosseye, Janina	211
Goudin, Yoann	364
Gow, Ann	177
Gradoux, Xavier	216
Graham, Wayne	502
Greenspan, Brian	259
Gregory, Ian	14, 179
Grimaldi, Kerri	529
Gross, Karin	559
Grossner, Karl	181
Grube, Nikolai	165
Grue, Dustin	183

Grumbach, Elizabeth	129, 130, 184, 213, 473, 498
Gschwind, Rudolf	159
Guan-tao, Jin	404
Guiliano, Jennifer	274
Gupta, Anshul	130
Gutierrez-Osuna, Ricardo	130
Gutiérrez, Silvia	288
Guzmán, Ana María	306
Hajj-Ahmad, Adi	44
Hamilton, Rachael	424
Hammond, Matthew	111
Han, Myung-Ja	465, 552
Handelman, Matthew	474
Hankinson, Andrew	186
Hanrahan, Elise	188
Hardie, Andrew	179
Harloff, Erik	407
Harrell, D. Fox	189
Hashimoto, Takako	191
Hashimoto, Yuta	475
Hasid, Tomer	451
Hautli-Janisz, Annette	117
Havemann, Leo	517
Hawker, Alex	115
Hayashi, Susumu	475
Heil, Jacob	498
Heitman, Carrie C.	476
Hennebert, Jérôme	193
Henrich, Andreas	383
Herbert, Bruce	515
Hettel, Jacqueline	15
Heuser, Ryan	70, 72
Hidalgo Urbaneja, María Isabel	195
Hinrichs, Erhard	196
Hinrichs, Uta	198
Hjartarson, Paul	27, 518
Hofmann, Beate	457
Holzinger, Katharina	117
Hoogerwerf, Maarten	391
Hoover, David L.	200, 202, 305
Hopwood, Elizabeth	477
Hori, Masahiro	477
Houghton, Hugh	204
Hours, Bernard	468
Houston, Natalie	206
Hove, Ingrid	231
Hsiang, Jieh	207
Hsu, Su-Chu	444, 449
Hughes, Lorna	4, 11, 209
Huistra, Hieke	210
Huitfeldt, Claus	360
Hunter, Jane	211
INKE-MVP Research Team	126
INKE Research Group,	440
Ichimura, Taro	225
Imahayashi, Osamu	477
Inoue, Satoshi	558
Isaksen, Leif	355
Ives, Maura	213
Jacke, Janina	264
Jackson, Cornell Alexander	215
Jacquin, Jérôme	216
Jakacki, Diane	217, 478
Jannidis, Fotis	64, 135
Jautze, Kim Johanna	480
Jiménez Mavillard, Antonio	120
Jänicke, Stefan	220
Jockers, Matthew	17
Joe, Bailey	308
Johnson, Ian R.	482
Johnsrud, Brian	405
Johnston, Penny	483
Joshgun, Sirajzade	443
Jui-sung, Yang	404
Juola, Patrick	219
Kahn, Andrew	448
Kao, Dominic	189
Kaplan, Frédéric	v, 222, 280, 330
Karadkar, Unmil	533
Katelnikoff, Joel	224
Katharina, Lorenz	308
Kathrin, Nühlen	443
Kawase, Akihiro	225
Kay, Christian	68
Keating, John	141, 315
Kee, Kevin Bradley	227
Keim, Daniel A.	117
Kejriwal, Gaurav	229
Kelland, Katharine Louise	295
Keller, Alice	18
Keller, Stefan Andreas	18
Kelley, Wyn	463
Kemman, Max	24, 488
Kenny, Julia	501
Kermes, Hannah	461
Kestemont, Mike	64, 135
Khan, Anas Fahad	109
Kilchenmann, André	366
Kinnaman, Alex	425
Kirilloff, Gabrielle	144
Kitamoto, Asanobu	484
Klaussner, Carmen	485
Klein, Benjamin Eliot	486
Klein, Lauren F.	154
Kleinman, Scott	327
Kleppe, Martijn	24, 488
Knauth, Jürgen	457
Knechtel, Ruth	439, 441
Koguchi, Keisuke	477
Kokkinakis, Dimitrios	497
Kolatzek, Robert	535
Kolly, Marie-José	231
Konstantelos, Leo	177
Koolen, Corina	233
Kraus, Kari	44, 241
Krause, Thomas	489
Krauwer, Steven	196
Kraxenberger, Maria	72
Kronenwett, Simone	523
Krones, Tim	535
Kösser, Sylwia	469
Küster, Marc Wilhelm	169
Kukita, Minao	475
Kulasalu, Kaisa	345
Kumari, Ashanka	490
Kurtz, Wendy	253
Lallemand, Carine	407
Lamé, Marion	109
Lamarra, Antonio	235
Lana, Maurizio	492
Lang, Anouk	14, 238, 494
Lange, Felix	389
Lavrentiev, Alexei	239
Law, Anita	70
Lawless, Séamus	419
Lawrence, Katharine Faith	19, 20
Lawton, Courtney	490
Lazzaro, Marilena	407
Lüdeling, Anke	489
Leblanc, Jean-Marc	240
Leemann, Adrian	231
Lein, Julie Gonnering	446
Leipold, Aletta	469
Leonard, Peter	382, 497
Lewis, Vivian	535
Liao, Wen-Hung	242
Lim, Chong-U	189
Lin, Cheng-Wei	444
Lin, Jimmy	241
Lindblad, Purdom	15
Lingold, Mary Caton	44
Lior, Wolf	486
Lipshin, Jason	189
Liu, Jyi-Shane	242

Llewellyn, Tanya	70
Lobin, Henning	95
Locuratolo, Elvira	378
Lombardi, Thomas	246
Long, Christopher	495
Lorang, Elizabeth M	248
Losh, Elizabeth	250
Lunde, Joseph	248
Lynch, John	253
MacCall, Steven L.	497
MacDonald, Andrew	439, 441
Macarthur, John	211
Macnamara, Craig	211
Magro, Diego	492
Mahony, Simon	402
Malm, Mats Ulrik	497
Mandell, Laura	184, 213, 416, 473
Mandell, Laura C.	498
Manovich, Lev	250
Marchand-Maillet, Stephane	165
Marchetti, Alessandro	255
Marcoux, Yves	360
Mareike, Hoeckendorff	22
Marín Pina, M ^a Carmen	500
Marques de Matos, Debora	540
Marras, Cristina	257
Martin, Kim	259, 261
Martin, Meredith	499
Martin, Worthy	104, 149
Martínez Cantón, Clara Isabel	174
Martos Pérez, María Dolores	174, 500
Maslov, Alexey	416
Masotti, Raffaele	501
Mastrangelo, Simon	343
Mauro, Aaron	263
Maycock, Keith	315
Mazzei, Andrea	330
McCann, Paul	209
McClure, David	502
McCue, Carmen	490
McCurdy, Nina	446
McGrath, Jim	502
Meeks, Elijah	181
Meghini, Carlo	378
Meister, Jan Christoph	264
Melenhorst, Marc	407
Mellish, Chris	86
Meneses, Luis	460
Meyer, Miriah	446
Micheel, Isabel	407
Mike, Heffernan	308
Miller, Ben	266
Miller, Laura	15
Milligan, Ian	268
Mimno, David	271
Molitor, Paul	469
Molloy, Laura	177
Montague, John Joseph	272
Moreno, Jose Luis	490
Morioka, Tomohiko	409
Morrissey, Robert	334
Mueller, Daren	44
Muller, A. Charles	277
Munoz, Trevor	274
Munson, Matthew	106, 275
Muñoz, Trevor	528
Murakami, Uesaka	547
Ó Murchú, Tomás	419
Murphy, Orla	26
Murrieta-Flores, Patricia	179
Muys, Andrae	211
Mylonas, Elli	10
Nagasaki, Kiyonori	277, 477
Nagel, Dylan	391
Nahli, Ouafae	109
Napolin, Julie Beth	149
Nelson, Brent	439
Nerbonne, John	485
Neudecker, Clemens	146, 184
Neuroth, Heike	18
Niederée, Claudia	436
Nieves, Angel David	529
Nikaido, Yoshihiro	409
Nishimura, Yoko	484
Nishio, Miyuki	477
Noecker Jr, John	278
Noël, Geoffroy	298, 539
Norberg, Brian	143
Norrish, Jamie	550
Novak, Jasminko	407
Nowviskie, Bethany	502
Nucci, Francesco	407
Nuessli, Marc-Antoine	280
Nunes Barreiros, Patrício	503
Nyhan, Julianne	46
O'Connor, Alex	434
O'Donnell, Daniel Paul	33, 54
O'Sullivan, James	147, 160, 281, 478
Oard, Doug	44
Odebrecht, Carolin	284, 489
Odobez, Jean-Marc	165
Ogden, Mitchell Paul	506, 506
Ogiso, Toshinobu	225
Ohge, Christopher M.	507
Ohura, Makoto	475
Olive, Jennifer	266
Oliveira, Lídia	285
Olivieri, Ryan	229
Ordelman, Roeland	24
Orekhev, Boris	100
Organisciak, Peter	507
Ortega, Érika	120, 288
Ott, Tobias	509
Ott, Wilhelm	509
Paixão de Sousa, Maria Clara	306
Pak, Burak	290
Pallan, Carlos	165
Palm, Fredrik	26
Papaelias, Amy	327
Papaki, Eliza	433
Paquette-Bigras, Ève	292
Partzsch, Henriette	368
Pasini, Chiara	407
Pöckelmann, Marcus	313
Peña, Ernesto	296, 439
Peaker, Alicia	502
Pedraça, Samia	459
Perdue, Susan	20
Perkins, Jody	106
Peroni, Silvio	492
Peterson, Noah	513
Petris, Marco	7
Pett, Daniel Edward John	294, 295
Piao, Scott	68
Pierazzo, Elena	298
Pieters, Toine	210, 299
Pino-Díaz, José	301
Pinto, Caro	466
Plale, Beth	508
Plasek, Aaron	303, 305
Pöllitz, Christian	428
Pluta, Izabella	58
Poitras, Eric	227
Porter, Dot	458
Porter, J.D.	72
Potvin, Sarah	466, 514
Powell, Daniel	516
Prats Lopez, Montserrat	368
Priani, Ernesto	306
Priani Saisó, Ernesto	467
Pridal, Petr	517
Priego, Ernesto	517
Priestnall, Gary	308
Prom, Christopher	310

Pérès, Marie	240
Pugin, Laurent	186
Punzalan, Ricardo L. Punzalan	241
Pytlik Zillig, Brian	85
Pytlowany, Anna	312
Qing-feng , Liu	404
Quamen, Harvey	27, 518
Quan-Haase, Anabel	259, 261
Quigley, Aaron	198
Quinn, Deirdre	315
Quinsland, Kirk	520
Radtke, Nadja	428
Radzikowska, Milena	439
Rahtz, Sebastian	10, 20, 439
Rajan, Vinodh	317
Ray Murray, Padmini	521
Rayner, Samantha	402
Rayson, Paul	68, 179
Reed, Scott Brian	28
Reeve, Jonathan Pearce	319
Regattieri, Lorena	459
Rehberger, Dean	495, 522
Reside, Doug	320, 321, 323
Reyes-Garcia, Everardo	324
Rhody, Lisa	326, 466
Riddell, Allen	350
Ridge, Mia	327
Ries, Thorsten	329
Rio, Alice	111
Risam, Roopika	54
Ritter, Jörg	469
Ritter, Julia	314
Roberts, Owain	209
Robyn, Sullivan	308
Rocchio, Michael	253
Rochat, Yannick	330
Rochester, Eric	502
Rockwell, Geoffrey	34, 64, 272, 357, 441
Rodríguez-Ortega, Nuria	301, 331
Roe, Glenn	334
Roeder, Geoff G.	439
Rohrdantz, Christian	117
Rojansky, Shay	474
Roman Rangel, Edgar	165
Romanello, Matteo	62
Romary, Laurent	29, 489
Rosenthaler, Lukas	18, 159, 335, 366
Ross, Stephen	126
Rosselli Del Turco, Roberto	337
Rossi Rognoni, Gabriele	471
Rouse, Rebecca	520
Rowberry, Simon	522
Roxin, Ioan	75
Rubin-Detlev, Kelsey	448
Ruecker, Stan	272, 439
Rupp, C.J.	179
Rusinek, Sinai	474
Rybicki, Jan	135, 281
Sack, Graham Alexander	339, 340
Sahle, Patrick	523
Saldana, Marie	342
Salzbrunn, Monika	343
Samuelson, Todd	129, 185
Sanz, Amelia	368
Sarv, Mari	345
Sayers, Jentery	33, 44, 126
Schöch, Christof	106, 135, 350
Scheirer, Walter	524
Schell, Justin	346
Scheuermann, Gerik	220
Schiller, Ines	95
Schirmacher, Peter	489
Schmunk, Stefan	9, 526
Schneider, Gerlinde	527
Schnepper, Rachel	12
Schoenberger, Zachary	459
Scholger, Walter	31, 347
Schreibman, Susan	348
Schroeder, Caroline T.	349
Schluthess, Sara	445
Schwartz, Frithjof	6
Schweizer, Tobias	366
Schwinghammer, Ylva	527
Seláf, Levente	174
Selig, Thomas	169
Sellmer, Megan	441
Semlak, Martina	352
Sensenbaugh, Jonny	72
Senseney, Megan	528
Seuffert, Janette	42
Shimoda, Masahiro	277
Shirazi, Roxanne	466
Shirota, Yukari	191
Shrestha, Ayush	266
Shweta, Roni	89
Siemens, Lynne	32, 353
Sievers, Martin	169, 204
Sikes, Sara	529
Sillaume, Ghislain	407
Simon, Rainer	355
Simons, Janet Thomas	529
Simpson, John	272
Simpson, John Edward	32
Sinclair, Stéfan	34, 134, 272, 356, 439, 455
Smith, Catherine	204
Smith, David	358
Smith, Dustin	533
Smith, Kathleen	526
Soh, Leen-Kiat	248
Solomon, Rory	163
Spence, Paul	54, 306
Sperberg-McQueen, Michael	35, 360
Spiro, Lisa	535
Sporleder, Caroline	535
Sprugnoli, Rachele	255
Söring, Sibylle	544
Süsstrunk, Sabine	83
Stack, Padraig	15
Stahmer, Carl	362
Stein , Christian	537
Steiner, Elisabeth	539
Stern Cahoy, Ellysa	76
Stokes, Peter A.	539
Stoltz, Michael	541
Stommel, Jesse	362
Storrer, Angelika	428
Stoyanova, Simona	5, 433
Streiter, Oliver	364
Strötgen, Jannik	7
Su, Hui	44
Subotic, Ivan	366
Suciù, Radu	542
Sula, Chris Alen	543
Sulger, Sebastian	117
Sullivan, Brenton	368
Sun, Iris	154
Suárez, Juan Luis	120, 413
Sutherland, Ainsley	189
Suzan, van Dijk	368
Suzuki, Takafumi	370
Swafford, Joanna	372
Sweeny, Robert C.H.	374
Syd, Bauman	10
Ta-Shma, Amnon	451
Tabata, Tomoji	375, 477
Tackett, Justin	72
Tagliasacchi, Marco	407
Tanherlini, Timothy R.	271
Tanigawa, Katie	126
Tardella, Michela	235
Tarpley, Bryan	129
Tarte, Segolene	46
Tasovac, Toma	31
Tavoni, Mirko	378

Taylor, Toniesha	150
Tcheng, David	44
Team, The INKE	402
Tedrow, Kimberly Ann	425
Teehan, Aja	141
Teich, Elke	461
Terras, Melissa	vi
Teufel, Isolde	535
Teule, Herman	445
Thaller, Manfred	11
Theibault, John Christopher	379
Thieberger, Nick	381
Thijssen, Michiel	391
Thiruvathukal, George Kuriakose	107
Thély, Nicholas	26
Thomann, Johannes	545
Thomas, Grace	248, 490
Tilton, Lauren	382
Tobias, Gradi	382
Tomabechi, Toru	277
Tomasek, Kathryn	385
Tomasi, Francesca	492
Tonelli, Sara	255
Tonra, Justin	546
Torabi, Katayoun	129
Toth, Michael	386
Trachsel, Alexandra	62
Tran, Nam Khanh	436
Trettien, Whitney	44
Trippel, Thorsten	37
Tullos, Allen E.	388
Tupman, Charlotte	5
Turkel, William J.	44
Uesaka, Ayaka	547
Underwood, Ted	64, 549
Unold, Martin	389
Van Zundert, Joris	63
Van den Heuvel, Charles	391
Van der Plaat, Deborah	211
Vasold, Gunter	393
Veentjer, Ubbo	544
Verbeke, Demmy	395
Verbeke, Johan	290
Verheul, Jaap	299
Verhoeven, Deb	115, 396
Vernus, Pierre	468
Versienti, Loredana	378
Vertan, Cristina	29, 397
Vieira, Miguel	550
Vincenzo, Croce	407
Vitale, Valeria	22, 551
Vitali, Fabio	492
Vitali Rosati, Marcello	455
Vogeler, Georg	398
Vosyliute, Ingrida	400
Wade, Mara R	552
Walkowski, Niels-Oliver	401
Wallace, Claire	86
Wang, Xuemao	535
Wangchuk, Dorji	486
Warren, Chandler	144
Warwick, Claire	402
Wassmer, Andreas	159
Watson, Matilda	540
Webster, Gemma	86
Wehrwein, James	144
Wei-Yun, Chiu	404
Wendrich, Willeke	553
Wenger, Alexandre	542
Wen-huei, Cheng	404
Westcott, Stephanie	36
Westermeier , Carola	553
Wettlaufer, Jörg	554
Widner, Michael	405
Wiede, Wiebke	436
Wieneke, Lars	407
Willis, Craig	465
Wilson, Meagan	499
Windsor, Jennifer	439, 459
Wittern, Christian	408, 409
Wolf, Lior	89
Wolff, Mark	555
Wolfgang, Lukas	443
Wortmann, Sven	457
Wu, Daniel	339
Wu, Min	44
Wubben, Sander	233
Wulfman, Clifford E.	50
Wygoda, Ynon	474
Wynne, Martin	37
Wythoff, Grant	499
Yamada, Taizo	558
Yamashita, Natsumi	370
Yamazaki, Naoki	409
Yasuoka, Koichi	409
Zastrow, Thomas	559
Zeldes, Amir	349
Zerr, Sergej	436
Zhitomirsky-Geffet, Maayan	411
Zielke, Dennis	489
Zusman, Benji	339
Zwicker, Heather	459
de Jong, Franciska	24
de Jong, Hayco	156
de Soto, Pau	355
deWaard, Andrew	415
de la Rosa, Javier	38, 413
duPlessis, Anton Raymund	416
Çöltekin, Çağrı	485
van Cranenburgh, Andreas	233
van Dalen-Oskam, Karina	156

